Content suppresses style: dimensionality collapse in contrastive learning

Evgenia Rusak[*1], **Patrik Reizinger**[*1], **Roland S. Zimmermann**[*1], **Oliver Bringmann**[1], **and Wieland Brendel**[1,2]

[1]University of Tübingen, Germany
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

Contrastive learning is a highly successful yet simple self-supervised learning technique that minimizes the representational distance of similar (positive) while maximizing it for dissimilar (negative) samples. Despite its success, our theoretical understanding of contrastive learning is still incomplete. Most importantly, it is unclear why the inferred representation faces a dimensionality collapse after SimCLR training and why downstream performance improves by removing the feature encoder's last layers (projector). We show that collapse might be induced by an inductive bias of the InfoNCE loss for features that vary little within a positive pair (content) while suppressing more strongly-varying features (style). When at least one content variable is present, we prove that a low-rank projector reduces downstream task performance while simultaneously minimizing the InfoNCE objective. This result elucidates a potential reason why removing the projector could lead to better downstream performance. Subsequently, we propose a simple strategy leveraging adaptive temperature factors in the loss to equalize content and style latents, mitigating dimensionality collapse. Finally, we validate our theoretical findings on controlled synthetic data and natural images.

1 Introduction

Self-Supervised Learning (SSL) enables the exploitation of large, unlabeled data sets by defining surrogate objectives in place of labels. Contrastive Learning (CL) belongs to SSL-methods relying on data augmentation to define positive (similar) and generally negative (dissimilar) samples [13, 3, 6, 2, 4]. Some SSL methods circumvent the need for negative samples [5, 7, 21].

Dimensionality collapse occurs when the representations do not capture some latent factors. In CL, it was first described by Chen et al. [3] as information loss—quantified by linear readout performance—occurring in the last layers (often called projector or projection head). The linear readout layer's weight matrix was found to be approximately low-rank, which led to the conjecture that the SimCLR loss causes information loss [3]. As a result, discarding the projector after training became popular in both contrastive [6, 2, 4] and non-contrastive methods [5, 7, 21].

We investigate why the commonly-used InfoNCE loss [13] leads to information loss in the projector [3]. While a previous study [11] already investigated the role of the projector, we find their analysis incomplete (see Appx. E for more details). In contrast, we interpret the information loss in the context of identifiability for CL [22], learning dynamics [1, 17, 11, 18], and the content-style partitioning of latent factors [19]. Using these tools, we theoretically confirm the conjecture of Chen et al. [3] by proving (for both linear and nonlinear networks) that a content-style partitioning exploits an inductive bias of the InfoNCE loss, leading to collapse. The seeming controversy to the results

[*]Equal contribution. Correspondence to `first.lastname@uni-tuebingen.de`.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

of [22]—i.e., that CL can invert the Data Generating Process (DGP)—is resolved by showing that the inductive bias of the InfoNCE loss is only effective when there is at least one content variable. We propose a possible explanation for the projector's role: by localizing dimensionality collapse, it compensates for the inductive bias of the InfoNCE loss, which can be thought of as shielding the intermediate representations. Our contributions are:

- We prove that the contrastive (SimCLR) loss prefers dimensionality collapse when the latent conditional distribution of positive pairs has invariant (content) and varying (style) components—for linear and *nonlinear* models, box and sphere latent spaces.
- We propose a strategy of adaptive temperature values in the InfoNCE loss to mitigate dimensionality collapse based on our insights. Further, we conjecture that the projector's role is to shield the intermediate representation by absorbing the low-rank transformations leading to collapse. Finally, we provide empirical verification on controlled synthetic data and natural images.

2 Background

Self-supervised/Contrastive learning. Contrastive Learning (CL) is a SSL paradigm learns an encoder $\boldsymbol{f} : \mathcal{X} \to \mathcal{Z}$ mapping observations $\boldsymbol{x}$ to latent vectors $\boldsymbol{z}$. CL uses positive pairs $\tilde{\boldsymbol{x}}$ (similar data point, e.g., augmentations of the same sample), and negative pairs $\boldsymbol{x}^-$ (dissimilar data points, e.g., uniformly drawn from the dataset). Describing observations by a DGP $\boldsymbol{g} : \mathcal{Z} \to \mathcal{X}$, the positive samples follow a conditional ($p\left(\tilde{\boldsymbol{z}}|\boldsymbol{z}\right)$), the unconditional samples a marginal $p\left(\boldsymbol{z}\right)$ distribution over $\mathcal{Z}$ [22]. Under certain assumptions, CL inverts the DGP, i.e., the composition $\boldsymbol{h} = \boldsymbol{f} \circ \boldsymbol{g}$ is a trivial map (including scaling, permutation, or affine transformations, depending on the assumptions) [22]. CL mostly uses losses from the InfoNCE family [8, 13] with the *normalized* form of:

$$\lim_{M\to\infty} \mathcal{L}_{\mathrm{CL}} - \log M + \log|\mathcal{Z}| = \mathbb{E}_{(\boldsymbol{x},\tilde{\boldsymbol{x}},\boldsymbol{x}^-)}\left[-\log \frac{e^{\boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{f}(\tilde{\boldsymbol{x}})/\boldsymbol{\tau}}}{e^{\boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{f}(\tilde{\boldsymbol{x}})/\boldsymbol{\tau}} + \sum_i^M e^{\boldsymbol{f}^T(\boldsymbol{x}^-)\boldsymbol{f}(\tilde{\boldsymbol{x}})/\boldsymbol{\tau}}}\right], \quad (1)$$

where M is the number of negative samples, $\boldsymbol{\tau}$ the scalar temperature, $|\mathcal{Z}|$ the normalization constant of the latents, $\tilde{\boldsymbol{x}} \sim p(\tilde{\boldsymbol{x}}|\boldsymbol{x})$ and $(\boldsymbol{x}, \boldsymbol{x}^-) \sim^{i.i.d.} p(\boldsymbol{x})$ the positive and negative observation pairs.

Content-style partitioning of $\mathcal{Z}$. von Kügelgen et al. [19] discuss a partitioning of $\boldsymbol{z} \in \mathcal{Z}$ into invariant (*content*) $\boldsymbol{z}^c \in \mathcal{Z}^c$ and changing (*style*) $\boldsymbol{z}^s \in \mathcal{Z}^s$ variables, where $\boldsymbol{z}^c$ of positive pairs follows a δ-distribution [19, Assum. 3.1] whereas the distribution of $\boldsymbol{z}^s$ of those pairs is not degenerate [19, Assum. 3.2], yielding $p\left(\tilde{\boldsymbol{z}}|\boldsymbol{z}\right) = \delta\left(\tilde{\boldsymbol{z}^c} - \boldsymbol{z}^c\right) p(\tilde{\boldsymbol{z}}^s|\boldsymbol{z}^s)$, where $\mathcal{Z} = \mathcal{Z}^c \times \mathcal{Z}^s$. We note that Jing et al. [11] implicitly use a similar distinction—defined as different augmentation amplitudes. Thus, their approach is akin to a generalized notion of content variables with $\boldsymbol{\sigma}_s^2 \gg \boldsymbol{\sigma}_c^2 > 0$.

Learning dynamics and collapse. Theoretical and empirical results [17, 11, 19] show that data properties affect learning dynamics of the latent components z_i; there seems to be a connection between the Hessian's rank and the number of classes [15, 16], based on which Awasthi et al. [1] show that the optimal representations for InfoNCE are Simplex Equiangular Tight Frames (ETFs) [14].

3 Theory

We provide theoretical insights on why dimensionality collapse possibly happens and a strategy to alleviate it. Our approach is investigating the loss, akin to [1], and showing that collapsing style variables reduces the loss compared to a representation capturing all latents (e.g., when $\boldsymbol{f} = \boldsymbol{g}^{-1}$). We consider linear and nonlinear models, and corresponding encoders with normalized (i.e., $\mathcal{S}^{d-1}$) and non-normalized latent spaces (e.g., $\mathbb{R}^d$), connecting insights from [22, 19, 1, 11]. We defer most formal results and proofs to Appx. D; we also include a detailed analysis of the linear case in Appx. C.

Assuming at least one content dimension, a linear DGP $\boldsymbol{g}$ and linear encoder weights $\mathbf{W}$; the question is for which $\mathbf{W}$ $\mathcal{L}_{\mathrm{CL}}$ is minimal. I.e., in the empirical loss formulation eq. (4), the norms for positive and negative pairs should change in a way that decreases the loss. To show that a low-rank $\mathbf{W}$ yields the optimum (Prop. 4), we leverage that $\boldsymbol{z}^c$ has a delta conditional: i.e., it does not contribute to the alignment term, so the norm w.r.t. the positive pairs $(\boldsymbol{z}, \tilde{\boldsymbol{z}})$ simplifies to:

$$\|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\|^2 = \left(\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\right)^T \mathbf{W}^T\mathbf{W}\left(\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\right) = \left(\boldsymbol{x}_i^s - \tilde{\boldsymbol{x}}_i^s\right)^T \mathbf{W}_s^T\mathbf{W}_s\left(\boldsymbol{x}_i^s - \tilde{\boldsymbol{x}}_i^s\right), \quad (2)$$

which is a weighted inner product with $\mathbf{W}_s$ affecting only $\boldsymbol{z}^s$. The larger $\boldsymbol{\sigma}_s^2$, the larger $(\boldsymbol{x}_i^s - \tilde{\boldsymbol{x}}_i^s)$ (on average). To reduce the norm, $\mathbf{W}$ should be such that the largest components of observation covariance belong to its kernel (i.e., they are collapsed)—this is possible when $\mathbf{W}$ becomes low-rank

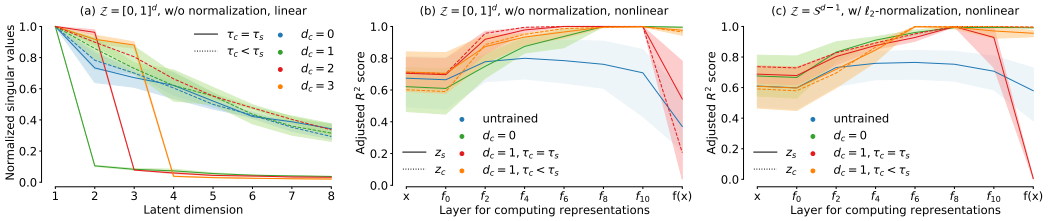

Figure 1: **Left:** Normalized singular values of trained weights of linear models for different temperatures: $\boldsymbol{\tau}_c = \boldsymbol{\tau}_s = \boldsymbol{\tau} = 1$ for all dimensions (solid) and $\boldsymbol{\tau}_c < \boldsymbol{\tau}_s$ (dashed). Setting $\boldsymbol{\tau}_c < \boldsymbol{\tau}_s$ yields the same spectrum as for the uncollapsed setting (i.e., $d_c = 0$). Adjusted R^2 score of style (solid) and content (dashed) dimensions when evaluated on intermediate layers throughout the encoder f. For both hyperbox (**middle**) and hypersphere (**right**), we observe collapse in the final layer when $\boldsymbol{\tau}_c = \boldsymbol{\tau}_s$, corroborating our linear model results. When setting $\boldsymbol{\tau}_c < \boldsymbol{\tau}_s$, the collapse is alleviated.

(with implications for its gradient discussed in Prop. 2). Note that this reasoning requires the presence of content variables: If only style variables are present, then a low-rank $\mathbf{W}$ will increase the norm for $(\boldsymbol{z}, \tilde{\boldsymbol{z}})$—intuitively, this is why $\mathbf{W} = \mathbf{0}$ does not minimize $\mathcal{L}_{\text{CL}}$, which we show in Appx. C.

Perhaps surprisingly, the linear results transfer to nonlinear DGPs $\boldsymbol{g}$ and encoders $\boldsymbol{f}$: making $\boldsymbol{f}$ low-rank can drive $\|\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{f}(\tilde{\boldsymbol{x}}_i)\|^2$ to zero, e.g., by scaling $\boldsymbol{f}$ with a low-rank diagonal matrix $\mathbf{D}$ ($\operatorname{rank}(\mathbf{D}) = d_c$). By the same argument as above, this results in a low-rank (collapsed) $\boldsymbol{f}$ that minimizes $\mathcal{L}_{\text{CL}}$ while having low-rank gradients (Prop. 5). We state our most general result for *unnormalized* $\boldsymbol{f}$ as (Im denoting the image, Ker the kernel):

Proposition 1. *[$\boldsymbol{z}^s$-collapse achieves optimal $\mathcal{L}_{\text{CL}}$] When $\operatorname{Im}(\boldsymbol{f}(\mathcal{X})) = \mathcal{Z}^c$ and $\operatorname{Ker}(\boldsymbol{f}) = \mathcal{Z}^s$ then $\mathcal{L}_{\text{CL}}$ achieves its optimum under the bijective encoder of Assum. 2.*

For normalized representations, we rely on [1], showing that for a DGP with class-determining latent variables, the optimal representations for the InfoNCE loss are Simplex ETFs [14], reducing the inter-class variance, thus, collapsing to the class means. We interpret the content variables $\boldsymbol{z}^c$ as the class-determining variables of the DGP, resulting in the following lemma:

Lemma 1. *[Interpreting $\boldsymbol{z}^c$ as class variables explain collapse on $\mathcal{S}^{d-1}$] When $\boldsymbol{z}^c$ are interpreted as the class variables of a DGP and [1, Assums. 3.1,3.6] hold then the optimal representation minimizing the InfoNCE loss yields collapsed $\boldsymbol{z}^s$.*

Our above results only show that there is an inductive bias in the InfoNCE loss that becomes relevant as soon as there is at least one content dimension. Fortunately, these results suggest a strategy for how we might avoid dimensionality collapse: by using different temperature values for content and style latents, we could counter the emerging low-rank structure in $\boldsymbol{f}$. We use $\boldsymbol{\tau}_c > \boldsymbol{\tau}_s$ (this is similar to the diagonal matrix $\mathbf{D}$ in our proofs) in the InfoNCE loss and demonstrate its success in § 4.

Engineering a collapsed but minimal-loss $\boldsymbol{f}$ also sheds light on a possible role of the projector: We conjecture that the selectivity of the low-rank diagonal matrix $\mathbf{D}$ materializes in the projector as well, concentrating dimensionality collapse in the last few layers. Thus, "shielding" the intermediate representations from further collapse; this is what we observe in practice (§ 4) and summarize as:

Conjecture 1. *The projector's role in CL is "shielding" the intermediate representation from the inductive bias of $\mathcal{L}_{\text{CL}}$.*

4 Experimental results

4.1 Fully controlled experiments

Preliminaries. In our first set of experiments, we follow Zimmermann et al. [22] and consider a fully controlled Data Generating Process (DGP), where the ground-truth latent space is either a hyperbox ($\mathcal{Z} = [0,1]^d$) or a hypersphere ($\mathcal{Z} = \mathcal{S}^{d-1}$). Specifically, we use an invertible Multi-Layer Perceptron (MLP) as the generative model $\boldsymbol{g}$ with same input, intermediate, and output dimensionality d (optionally, with Leaky ReLU non-linearities). Unless noted otherwise, $d = 10$. Further, the latents of anchor points and negative samples are sampled from a latent marginal distribution $\boldsymbol{z} \sim p(\boldsymbol{z})$, while positive samples follow a conditional distribution $\tilde{\boldsymbol{z}} \sim p(\tilde{\boldsymbol{z}}|\boldsymbol{z})$. During training, the model only

uses observations, which are obtained by passing latents $\boldsymbol{z}$ through $\boldsymbol{g}$. The encoder $\boldsymbol{f}$ is also modeled as an MLP with same input, intermediate and output dimensionality, optionally, with Leaky ReLU non-linearities. As we have access to the ground-truth $\boldsymbol{z}$, we compute the amount of information in the inferred latents $\boldsymbol{z}'$ about each of ground-truth latent dimensions by fitting a linear map $\mathbf{A}$ between $\boldsymbol{z}'$ and $\boldsymbol{z}$ and calculating the coefficient of determination R^2 [20] between $\mathbf{A}\boldsymbol{z}'$ and $\boldsymbol{z}$.

Linear DGP. We start with a linear DGP $\boldsymbol{g}$ and linear encoder $\boldsymbol{f}$ (without bias and $d = 8$ latent dimensions) to understand whether collapse happens and potentially how to alleviate it. The latent marginal is uniform, the conditional is normal with mean $\boldsymbol{z}$ for style dimensions and a δ-distribution for content dimensions. When $\mathcal{Z}$ is an ℓ_2-normalized hypersphere we train $\boldsymbol{f}$ using InfoNCE based on cosine-similarity; while for the hyperbox we use an extended InfoNCE objective for non-normalized representations based on an ℓ_p norm [22]. Inspecting the trained weight spectrum (varying d_c), we notice that $\boldsymbol{z}^s$ collapses when at least one content dimension is present, i.e., $d_c > 0$, unless we set smaller temperature for $\boldsymbol{z}^c$ (Fig. 1), corroborating our strategy Appx. C. Moreover, a collapsed $\boldsymbol{f}$ leads to a lower loss than an encoder that perfectly inverts the DGP, i.e., $\boldsymbol{f} = \boldsymbol{g}^{-1}$ (see Prop. 4).

Model	$\mathcal{Z} = [0,1]^d$			$\mathcal{Z} = \mathcal{S}^{d-1}$		
	$\mathcal{L}_{\text{align}}$	$\mathcal{L}_{\text{uniform}}$	$\mathcal{L}_{\text{CL}}$	$\mathcal{L}_{\text{align}}$	$\mathcal{L}_{\text{uniform}}$	$\mathcal{L}_{\text{CL}}$
$\boldsymbol{f} = \boldsymbol{g}^{-1}$	$0.02_{0.01}$	$7.69_{0.20}$	$7.71_{0.20}$	$-0.19_{0.32}$	$9.26_{0.00}$	$9.07_{0.32}$
$\boldsymbol{f}_{\text{InfoNCE}}$	$0.23_{0.14}$	$0.66_{0.08}$	$0.89_{0.11}$	$-0.97_{0.04}$	$9.29_{0.03}$	$8.32_{0.05}$

Table 1: With content dimensions present, collapsed have a lower loss than the intended solutions (i.e., $\boldsymbol{f} = \boldsymbol{g}^{-1}$). Loss was computed on 10'000 samples.

Nonlinear DGP. Next, we consider nonlinear DGPs (3-layer MLP, $d = 10$) and nonlinear encoders (7-layer MLP). Again, as latent spaces we consider a hyperbox as well as the hypersphere; the marginals are uniform, the conditionals of $\boldsymbol{z}^s$ are Gaussian (box) and von Mises-Fisher (vMF) (sphere) while that of $\boldsymbol{z}^c$ variables is a δ-distribution. We evaluate the adjusted R^2 score between the ground-truth latents and intermediate activations throughout the trained encoder $\boldsymbol{f}$, and observe that collapse is only present in the last layer (i.e., the projector) when $\boldsymbol{\tau}_c = \boldsymbol{\tau}_s$. With $\boldsymbol{\tau}_c < \boldsymbol{\tau}_s$, collapse is alleviated for the box, whereas for the sphere, we further need to remove the ℓ_2-normalization—a possible explanation is that optimization on $\mathcal{S}^{d-1}$ is harder; further, Papyan et al. [14] showed that collapsed representations form a Simplex ETF, i.e., they are already on $\mathcal{S}^{d-1}$ (Lemma 4).

Model	Lin. readout acc. after p.	before p.
Regular	$91.4_{0.5}$	$93.0_{0.1}$
w/o norm.	$91.4_{0.6}$	$92.5_{0.5}$
w/o norm + diff. $\boldsymbol{\tau}$	$92.5_{0.2}$	$93.2_{0.1}$

Table 2: Removing ℓ_2-normalization improves readout performance after and (slightly) before the projector.

4.2 Reducing dimensionality collapse on CIFAR10

On CIFAR10 [12] the ground-truth latents are unknown. Thus, we can only reason about recovered or collapsed latents indirectly: either by (1) characterizing the latent covariance spectrum after the projector [11], or by (2) training a linear readout for downstream classification [3]. We train a ResNet50 [9] on CIFAR10 [12] with the code from [10] and report the average of 3 random seeds. We find that removing the ℓ_2-normalization and standardizing the outputs substantially reduces dimensionality collapse after the projector (Fig. 2), but does not improve the linear readout accuracy (§ 4.1). Following the intuition that the distributions generating the latent factors in CIFAR10 have different variances, we set different $\boldsymbol{\tau}$ values for different dimensions. Specifically, we use $\boldsymbol{\tau} = 1$ for half of the dimensions and $\boldsymbol{\tau} = 0.7$ for the rest. Setting different $\boldsymbol{\tau}$ values does not influence the spectrum of the representations but improves the linear readout accuracy for representations both before and after the projector.

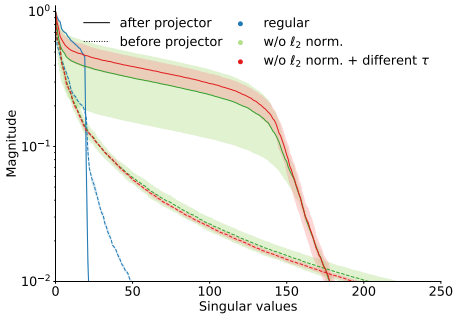

Figure 2: Spectrum of the representations' cov. after (solid) and before (dashed) the projector. Removing the ℓ_2-normalization reduces dimensionality collapse.

5 Conclusion

We propose a possible explanation for dimensionality collapse in CL by exposing a hidden inductive bias in the InfoNCE loss, which results in information loss of the style variables. Based on these insights, we provide a possible explanation for the relevance of a projector and suggest a strategy to alleviate this phenomenon, resulting in a reduced collapse on both synthetic data and on CIFAR10.

Acknowledgments and Disclosure of Funding

We thank Julian Bitterwolf and Steffen Schneider for valuable discussions.

This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A & 01IS18039B, and by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. Wieland Brendel acknowledges financial support via an Emmy Noether Grant funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Evgenia Rusak, Patrik Reizinger, Roland S. Zimmermann. Patrik Reizinger acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program.

References

[1] Pranjal Awasthi, Nishanth Dikkala, and Pritish Kamath. Do More Negative Samples Necessarily Hurt In Contrastive Learning? In *Proceedings of the 39th International Conference on Machine Learning*, pages 1101–1116. PMLR, June 2022. URL https://proceedings.mlr.press/v162/awasthi22b.html. ISSN: 2640-3498. 1, 2, 3, 12

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]*, June 2020. URL http://arxiv.org/abs/2002.05709. arXiv: 2002.05709. 1, 4

[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 1

[5] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning, November 2020. URL http://arxiv.org/abs/2011.10566. Number: arXiv:2011.10566 arXiv:2011.10566 [cs]. 1

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised Learning. *arXiv:2006.07733 [cs, stat]*, September 2020. URL http://arxiv.org/abs/2006.07733. arXiv: 2006.07733. 1

[8] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90. 4

[10] Tianyu Hua. A pytorch implementation for simsiam. https://github.com/PatrickHua/SimSiam, 2021. 4

[11] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding Dimensional Collapse in Contrastive Self-supervised Learning, April 2022. URL http://arxiv.org/abs/2110.09348. Number: arXiv:2110.09348 arXiv:2110.09348 [cs]. 1, 2, 4, 8, 9, 10, 13

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 4

[13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv: 1807.03748. 1, 2

[14] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, October 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015509117. Publisher: Proceedings of the National Academy of Sciences. 2, 3, 4, 8

[15] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the Hessian in Deep Learning: Singularity and Beyond, October 2017. URL http://arxiv.org/abs/1611.07476. arXiv:1611.07476 [cs]. 2

[16] Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical Analysis of the Hessian of Over-Parametrized Neural Networks, May 2018. URL http://arxiv.org/abs/1706.04454. arXiv:1706.04454 [cs]. 2

[17] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv:1312.6120 [cond-mat, q-bio, stat]*, February 2014. URL http://arxiv.org/abs/1312.6120. arXiv: 1312.6120. 1, 2

[18] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised Learning Dynamics without Contrastive Pairs, October 2021. URL http://arxiv.org/abs/2102.06810. arXiv:2102.06810 [cs]. 1

[19] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style, June 2021. URL http://arxiv.org/abs/2106.04619. arXiv: 2106.04619. 1, 2

[20] Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585, 1921. 4

[21] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *arXiv:2103.03230 [cs, q-bio]*, June 2021. URL http://arxiv.org/abs/2103.03230. arXiv: 2103.03230. 1

[22] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive Learning Inverts the Data Generating Process. *arXiv:2102.08850 [cs]*, February 2021. URL http://arxiv.org/abs/2102.08850. arXiv: 2102.08850. 1, 2, 3, 4

A Useful lemmas

Lemma 2 (Rank upper bound of matrix products)**.** *For matrices $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times k}$ the rank of the product $\mathbf{AB}$ is upper bounded by the minimum of the ranks of $\mathbf{A}$, $\mathbf{B}$, i.e.,*

$$\operatorname{rank}(\mathbf{AB}) \leq \min\left(\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\right). \tag{3}$$

Lemma 3 (Singularity of matrix product)**.** *For square matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$, the product $\mathbf{AB}$ is singular if and only if both $\mathbf{A}$, $\mathbf{B}$ are singular—including the case when any of both is the zero matrix $\mathbf{0} \in \mathbb{R}^{n\times n}$.*

Lemma 4 (Collapsed latent class means lies on $\mathcal{S}^{d-1}$ [14])**.** *For a nonlinear encoder $\boldsymbol{f}$ with representations $\boldsymbol{z}$ (used for classification via linear readout) the class means collapse to a Simplex ETF on $\mathcal{S}^{d-1}$.*

B Assumptions

Assumption 1 (Bijective linear DGP)**.** *We assume that for latents $\boldsymbol{z} \in \mathbb{R}^d$, the observations $\boldsymbol{x} \in \mathbb{R}^D$ are generated as $\boldsymbol{x} = \mathbf{W}^{-1}\boldsymbol{z}$. For the inverse to exists, we assume $d = D$.*

Assumption 2 (Bijective nonlinear DGP)**.** *We assume that for latents $\boldsymbol{z} \in \mathbb{R}^d$, the observations $\boldsymbol{x} \in \mathbb{R}^D$ are generated as $\boldsymbol{x} = \boldsymbol{g}\boldsymbol{z}$.*

Assumption 3 (Conditional and marginal axes align)**.** *The covariances of the marginal and conditional are simultaneously diagonalizable, i.e., the directions of the corresponding variances align. Furthermore, w.l.o.g., we assume that both conditional and marginal covariances are diagonal w.r.t. the canonical basis.*

Assumption 4 (Informative conditional)**.** *We assume that we have informative conditional distributions in the latent space, i.e., $\boldsymbol{\sigma}_c^2 \ll \boldsymbol{\sigma}_s^2 < \boldsymbol{\sigma}_z^2$, where $\boldsymbol{\sigma}_c^2$, $\boldsymbol{\sigma}_s^2$ and $\boldsymbol{\sigma}_z^2$ denote the variance of the positive pair's conditional distribution for the content, style and all dimensions, respectively.*

C Gradient dynamics in the linear case

For our analysis, we will rely on the empirical formulation of $\mathcal{L}_{\mathrm{CL}}$ (eq. (1)), i.e.:

$$\sum_i \sum_{j\neq i} -\log \frac{\exp\left(-\frac{\|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\|^2}{2}\right)}{\exp\left(-\frac{\|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\|^2}{2}\right) + \sum_{j\neq i}\exp\left(-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2}{2}\right)}, \tag{4}$$

which has a minimum of zero—for all exponential terms being non-negative and the numerator being less than or equal to the denominator. Note that w.l.o.g. we set $\tau = 1$, as the temperature value only scales all components of the gradient equally in the linear case; thus, the temperature is irrelevant for this analysis.

We follow [11] to analyze the gradient of $\mathcal{L}_{\mathrm{CL}}$ (eq. (1)) w.r.t. the learned weight matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$. The expression for the temporal dynamics for $\mathbf{W}$ yields ($\tilde{z}$ are the latents generating the positive pairs from the conditional) for a large number of negative samples M:

Lemma 5 ($\mathbf{W}$ dynamics)**.** *Under the linear model (Assum. 1) and with a large number of negative samples ($M \to \infty$), the temporal dynamics (i.e., the negative gradient) of the weight matrix $\mathbf{W}$ depends on the marginal ($\boldsymbol{\Sigma}_{\boldsymbol{x}}$) and conditional ($\boldsymbol{\Sigma}_{\tilde{\boldsymbol{x}}}$) data covariance matrices.*

$$\frac{\mathrm{d}\mathbf{W}}{\mathrm{d}t} = \mathbf{W}(M+1)\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\tilde{\boldsymbol{x}}}\right], \tag{5}$$

Proof. The total derivative of the loss is:

$$\frac{\mathrm{d}\mathcal{L}_{\mathrm{CL}}}{\mathrm{d}\mathbf{W}} = \sum_{\boldsymbol{z},\tilde{\boldsymbol{z}}} \frac{\partial \mathcal{L}_{\mathrm{CL}}}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \mathbf{W}}, \tag{6}$$

yielding

$$\frac{\partial \mathcal{L}_{\mathrm{CL}}}{\partial \boldsymbol{z}_i} = (\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i) - \alpha_{ii}\,(\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i) - \sum_{j\neq i} \alpha_{ij}\,(\boldsymbol{z}_i - \boldsymbol{z}_j) = \sum_{j\neq i} \alpha_{ij}\,[(\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i) - (\boldsymbol{z}_i - \boldsymbol{z}_j)]$$

$$\frac{\partial \mathcal{L}_{\mathrm{CL}}}{\partial \tilde{\boldsymbol{z}}_i} = -\,(\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i) + \alpha_{ii}\,(\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i) = -\sum_{j\neq i} \alpha_{ij}\,(\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i)$$

$$\alpha_{ij} = \frac{1}{Z_i} \exp\left(-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2}{2}\right)$$

$$\alpha_{ii} = \frac{1}{Z_i} \exp\left(-\frac{\|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\|^2}{2}\right)$$

$$\alpha_{ii} + \sum_{j\neq i} \alpha_{ij} = 1$$

$$Z_i = \exp\left(-\frac{\|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\|^2}{2}\right) + \sum_{j\neq i} \exp\left(-\frac{\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2}{2}\right)$$

$$\frac{\partial \mathcal{L}_{\mathrm{CL}}}{\partial \boldsymbol{z}} = \boldsymbol{x}^T$$

Following Jing et al. [11], we describe the **negative** gradient of $\mathbf{W}$:

$$\frac{\mathrm{d}\mathbf{W}}{\mathrm{d}t} = -\sum_i \sum_{j\neq i} \alpha_{ij} \left([(\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i) - (\boldsymbol{z}_i - \boldsymbol{z}_j)]\, \boldsymbol{x}_i^T - (\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i)\, \tilde{\boldsymbol{x}}_i^T \right) \tag{7}$$

$$= -\sum_i \sum_{j\neq i} \alpha_{ij} \left[(\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i)\,(\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i)^T - (\boldsymbol{z}_i - \boldsymbol{z}_j)\, \boldsymbol{x}_i^T \right] \tag{8}$$

$$= -\sum_i \left[\left(\sum_{j\neq i} \alpha_{ij} \right) (\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i)\,(\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i)^T - \sum_{j\neq i} \alpha_{ij}\,(\boldsymbol{z}_i - \boldsymbol{z}_j)\, \boldsymbol{x}_i^T \right] \tag{9}$$

We note that as $M \to \infty$, due to $\sum_j \alpha_{ij} = 1$:

$$\alpha_{ij} \to \frac{1}{M} \qquad\qquad \sum_{j\neq i} \alpha_{ij} \to 1.$$

Thus, its variance also goes to zero, yielding an estimate for the conditional data covariance, namely:

$$\sum_i \left[\left(\sum_{j\neq i} \alpha_{ij} \right) (\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i)\,(\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i)^T \right] \to \mathbf{W}\,(M+1)\,\boldsymbol{\Sigma}_{\tilde{\boldsymbol{x}}}. \tag{10}$$

Moreover, as marginal samples are independent and identically distributed (i.i.d.), the covariance of $\boldsymbol{x}_i \boldsymbol{x}_j$ is zero, resulting in:

$$-\sum_i \sum_{j\neq i} \alpha_{ij}\,(\boldsymbol{z}_i - \boldsymbol{z}_j)\, \boldsymbol{x}_i^T = -\mathbf{W} \sum_i \sum_{j\neq i} \alpha_{ij} \left(\boldsymbol{x}_i \boldsymbol{x}_i^T - \boldsymbol{x}_j \boldsymbol{x}_i^T \right) \tag{11}$$

$$= -\mathbf{W} \sum_i \sum_{j\neq i} \alpha_{ij} \boldsymbol{x}_i \boldsymbol{x}_i^T \tag{12}$$

$$\sim -\mathbf{W} \sum_i \sum_{j\neq i} \frac{1}{M} \boldsymbol{x}_i \boldsymbol{x}_i^T = -\mathbf{W}\,(M+1)\,\boldsymbol{\Sigma}_{\boldsymbol{x}}. \tag{13}$$

Assembling the terms yields the total derivative:

$$\frac{\mathrm{d}\mathbf{W}}{\mathrm{d}t} = -\mathbf{W}\,(M+1)\,[\boldsymbol{\Sigma}_{\tilde{\boldsymbol{x}}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}], \tag{14}$$

□

The assumption of sufficiently large number of negative samples (or, equivalently, batch size, i.e., $M + 1$) in Lemma 5 also pinpoints a possible mechanism why increasing M can help CL:

Lemma 6 (Batch size affects gradient rank.)**.** *Increasing the batch size (or, equivalently, M), improves the estimate of the covariances $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{x}}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}$, leading to better conditioned (higher-rank) gradients.*

A simple way to see this is that if $M + 1 < d$ then the covariance matrix estimates cannot not even be full rank, since there will fewer terms than dimensions (this can be thought as a spectral decomposition with less than d components.

The eigenvalues of $[\boldsymbol{\Sigma}_{\tilde{\boldsymbol{x}}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}]$ *can* affect whether there is collapse. To connect eq. (5) to content-style paritioning, we reformulate it in terms of the latent marginal and conditional covariances (using the abbreviation ground truth (GT)):

$$[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\tilde{\boldsymbol{x}}}] = \mathbf{W}_{\text{GT}}^{-1} [\boldsymbol{\Sigma}_{\boldsymbol{z}} - \boldsymbol{\Sigma}_{\tilde{\boldsymbol{z}}}] \mathbf{W}_{\text{GT}}^{-T} = \mathbf{W}_{\text{GT}}^{-1} \text{diag}\left(\boldsymbol{\sigma}_z^2 - \boldsymbol{\sigma}_c^2; \boldsymbol{\sigma}_z^2 - \boldsymbol{\sigma}_s^2\right) \mathbf{W}_{\text{GT}}^{-T} \tag{15}$$

Jing et al. [11] reasons that $[\boldsymbol{\Sigma}_{\tilde{\boldsymbol{x}}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}]$ will be at least indefinite (with at least one negative eigenvalue) for "strong" augmentations (on the data). Nonetheless, under the assumption that the positive pairs are drawn from *informative conditionals* (i.e., their variance is smaller than that of the marginal, cf. Assum. 4), $[\boldsymbol{\Sigma}_{\tilde{\boldsymbol{x}}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}]$ cannot be negative semi-definite (NSD), it will remain positive definite (PD). Notably, "strong augmentations" turn out to be a *sufficient, but not necessary condition* for dimensionality—even [11, Fig. 3] demonstrates this fact where collapse is observed for $k < 1$, i.e., for informative conditionals. Equation (15) also states that the gradient for each element in $\mathbf{W}$ will be proportional to its actual value (since $[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\tilde{\boldsymbol{x}}}]$ cannot be negative definite (ND)). The data covariance decomposition in eq. (15) elucidates how content-style partitioning affects the gradients:

Lemma 7 ($\boldsymbol{z}^c$ has larger gradients)**.** *Under Assums. 1 and 3, the gradients of $\mathbf{W}$ in linear CL are larger for dimensions encoding content variables.*

Proof. The decomposition of $[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\tilde{\boldsymbol{x}}}]$ in eq. (5) as in eq. (15) shows that since $\boldsymbol{\sigma}_c^2 \ll \boldsymbol{\sigma}_s^2$ (Assum. 4), the gradients for $\boldsymbol{z}^c$ (defined as the gradient directions corresponding to the singular values $\boldsymbol{\sigma}_z^2 - \boldsymbol{\sigma}_c^2$) are larger. □

Note that Lemma 7 does not require that $\boldsymbol{\sigma}_c^2 = 0$, only that $\boldsymbol{\sigma}_c^2 \ll \boldsymbol{\sigma}_s^2$. Thus, holds even when the standard definition of $\boldsymbol{z}^c$ (with zero conditional variance) is relaxed. We can use the above insight to determine d_c (note that this can differ based on the output dimensionality):

Proposition 2 (Gradient spectrum determines d_c)**.** *The component-wise singular values of the gradient spectrum eq.* (5) *can be used to determine d_c, more precisely, the number of latent dimensions that encode (mostly) content variables.*

Proof. The proof follows trivially from Lemma 7, since content dimensions have larger norms in the gradient. □

Furthermore, we can use this insight to propose a heuristic how to scale the different gradients such that they will have the same order of magnitude:

Proposition 3. *The temperature for $\boldsymbol{z}^c$ should be $\boldsymbol{\tau}_c \sim \sqrt{\alpha_c} = \boldsymbol{\sigma}_c / \boldsymbol{\sigma}_s$ if $\boldsymbol{\tau}_s = 1$.*

Proof. Our goal is to achieve a spectrum with appr. the same singular values. For this, we need:

$$\frac{\boldsymbol{\sigma}_z^2 - \boldsymbol{\sigma}_c^2/\alpha_c}{\boldsymbol{\sigma}_z^2 - \boldsymbol{\sigma}_s^2} \sim 1 \tag{16}$$

$$1 - \frac{\boldsymbol{\sigma}_c^2/\alpha_c}{\boldsymbol{\sigma}_z^2} \sim 1 - \frac{\boldsymbol{\sigma}_s^2}{\boldsymbol{\sigma}_z^2} \tag{17}$$

$$\frac{1}{\alpha_c} \sim \frac{\boldsymbol{\sigma}_s^2}{\boldsymbol{\sigma}_c^2} \tag{18}$$

If we use $\boldsymbol{\tau}$ to induce this equalizing effect, then $\boldsymbol{\tau} \sim \sqrt{\alpha_c} = \boldsymbol{\sigma}_c / \boldsymbol{\sigma}_s$ □

Consequences for the optimal loss. Under our assumptions, in the $M \to \infty$ regime, the temporal dynamics of $\mathbf{W}$ is given by eq. (5). For this to be the zero matrix $\mathbf{0}$, both matrix terms need to be singular (Lemma 3). This implies that–by Assum. 4–$\mathbf{W}$ needs to be $\mathbf{0}$. However, in this case $\mathcal{L}_{\mathrm{CL}}$ is not minimal. Using the linear model from Assum. 1—i.e., $\boldsymbol{z} = \mathbf{W}\boldsymbol{x}$—, we calculate $\mathcal{L}_{\mathrm{CL}}$ for $\mathbf{W} = \mathbf{0}$:

$$\mathcal{L}_{\mathrm{CL0}} = \sum_i \sum_{j \neq i} -\log \frac{1}{1+M} = (M+1)\, M \log\left(1+M\right),$$

which is clearly suboptimal. Thus, collapsing all dimensions is not a preferable solution. In the following, we will point out an *inductive bias* in $\mathcal{L}_{\mathrm{CL}}$ that shows that i) a collapsed solution achieves the optimum of eq. (4); and ii) the collapsed solution has a lower $\mathcal{L}_{\mathrm{CL}}$ than the optimal solution, which we define as the inverse of the GT matrix $\mathbf{W}_{\mathrm{GT}}^{-1}$.

D Proofs

D.1 Proof of Prop. 4

Proposition 4 ($\boldsymbol{z}^s$-collapse achieves optimal $\mathcal{L}_{\mathrm{CL}}$)**.** *When* $\mathrm{Im}\left(\mathbf{W}\mathcal{X}\right) = \mathcal{Z}^c$ *and* $\mathrm{Ker}\left(\mathbf{W}\right) = \mathcal{Z}^s$ *then* $\mathcal{L}_{\mathrm{CL}}$ *achieves its optimum under the linear model of Assum. 1*

Proof. Assuming $\boldsymbol{z}^s$ collapse means that $\mathrm{rank}\left(\mathbf{W}\right) = d_c$ and that the norm containing the positive pairs, $\left\|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}_i\right\|^2$, in eq. (4) is zero, yielding:

$$\sum_i \sum_{j \neq i} -\log \frac{1}{1 + \sum_{j \neq i} \exp\left(-\frac{\left\|\boldsymbol{z}_i - \boldsymbol{z}_j\right\|^2}{2}\right)}. \tag{19}$$

Rewriting the norm as a weighted inner product results in:

$$\left\|\boldsymbol{z}_i - \boldsymbol{z}_j\right\|^2 = \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^T \mathbf{W}^T \mathbf{W} \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right). \tag{20}$$

Now we show that there is a construction of the inferred weight $\mathbf{W}$ such that the loss goes to zero, achieving its optimum. For this, we investigate the degrees of freedom of manipulating $\mathbf{W}$ while maintaining dimensionality collapse for $\boldsymbol{z}^s$—however, for our reasoning, it will be sufficient to show that there is a single construction admissible by these degrees of freedom. Scaling and rotation does not alter the inferred covariance of the latent space (i.e., we could consider a new matrix in the form of $\mathbf{W} \leftarrow \mathbf{RDW}$, where $\mathbf{R}$ is a rotation and $\mathbf{D}$ a diagonal scaling matrix). We will consider scaling by a diagonal matrix $\mathbf{D}$ and show that there exist a $\mathbf{D}$ such that $\mathcal{L}_{\mathrm{CL}}$ goes to zero. Namely, select a new $\mathbf{W} \leftarrow \mathbf{DW}$ such that at least $\left(\mathbf{D}\right)_{ii} \to \infty : \forall i \in \{1, \ldots, d_c\}$ (as $\mathbf{W}$ has rank d_c, the diagonal values corresponding to the style dimensions do not matter). In this case, as the term $-\left\|\boldsymbol{z}_i - \boldsymbol{z}_j\right\|^2$ goes to minus infinity, the exponential goes to 0 and, finally, yields $\mathcal{L}_{\mathrm{CL}} = 0$, concluding the proof. □

It remains to be seen whether the optimal $\mathbf{W}$ (i.e., the inverse of the ground-truth DGP matrix, denoted by $\mathbf{W}_{\mathrm{GT}}$) can have a lower loss value. We show that dimensionality collapse can decrease the loss value (note that rotation will not change the loss value):

Corollary 1 ($\mathbf{W}_{\mathrm{GT}}$ suboptimal when $d_c > 0$)**.** *The optimum of* $\mathcal{L}_{\mathrm{CL}}$ *w.r.t.* $\mathbf{W}$ *is not its true value when* $d_c > 0$*.*

Indirect. Assume that the true $\mathbf{W}$ is the optimum. Scale $\mathbf{W}$ by a diagonal matrix $\mathbf{D}$ such that it has only nonzero elements for the content dimensions—the presence of content dimensions is crucial; otherwise, $\mathbf{D}$ would affect the alignment term. This does not affect the alignment term but will increase $\left\|\boldsymbol{z}_i - \boldsymbol{z}_j\right\|^2$. Since the loss has form $-\log\left(a/\left(a+b\right)\right)$ and b will go to zero, the loss decreases. Thus, by contradiction, our proposition holds. □

D.2 Proof of Prop. 1

Proposition 1. *[*$\boldsymbol{z}^s$*-collapse achieves optimal* $\mathcal{L}_{\mathrm{CL}}$*] When* $\mathrm{Im}\left(\boldsymbol{f}\left(\mathcal{X}\right)\right) = \mathcal{Z}^c$ *and* $\mathrm{Ker}\left(\boldsymbol{f}\right) = \mathcal{Z}^s$ *then* $\mathcal{L}_{\mathrm{CL}}$ *achieves its optimum under the bijective encoder of Assum. 2.*

Proof. We apply the same strategy as in the proof of Prop. 4 (Appx. D.1), i.e., using a low-rank diagonal matrix $\mathbf{D}$ to compose a new encoder $\mathbf{D} \circ \boldsymbol{f}$ □

D.3 Proof of Prop. 5

Proposition 5 (The gradient rank at the optimum is at most d_c if $d_c > 0$)**.** *When there is at least one content variable, i.e., $d_c > 0$, the gradient of the loss has rank of at most d_c at the optimum.*

Proof. From Prop. 1 we know that if $\mathcal{L}_{\text{CL}}$ is minimal, a $\mathbf{D} \circ \boldsymbol{f}$ with a low-rank diagonal $\mathbf{D}$ is a possible encoder structure. Expressing the gradient of $\mathcal{L}_{\text{CL}}$ at the optimum w.r.t. a weight matrix $\mathbf{W}$ for any $\boldsymbol{z}$ (can be positive or negative pair as well), where we denote the intermediate representation after $\boldsymbol{f}$ but before $\mathbf{D}$ as $\boldsymbol{z}$ and after $\mathbf{D}$ as $\boldsymbol{z}_{\mathbf{D}}$:

$$\frac{\mathrm{d}\mathcal{L}_{\text{CL}}}{\mathrm{d}\mathbf{W}} = \sum_{\boldsymbol{z},\tilde{\boldsymbol{z}}} \frac{\partial \mathcal{L}_{\text{CL}}}{\partial \boldsymbol{z}_{\mathbf{D}}} \frac{\partial \boldsymbol{z}_{\mathbf{D}}}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{f}} \frac{\partial \boldsymbol{f}}{\partial \mathbf{W}}. \tag{21}$$

The second item in the chain rule yields

$$\frac{\partial \boldsymbol{z}_{\mathbf{D}}}{\partial \boldsymbol{z}} = \mathbf{D}, \tag{22}$$

which is constant for all latents and rank-deficient. Thus, the gradient can have a rank at most d_c by Lemma 2. □

D.4 Proof of Lemma 1

Lemma 1. *[Interpreting $\boldsymbol{z}^c$ as class variables explain collapse on $\mathcal{S}^{d-1}$] When $\boldsymbol{z}^c$ are interpreted as the class variables of a DGP and [1, Assums. 3.1,3.6] hold then the optimal representation minimizing the InfoNCE loss yields collapsed $\boldsymbol{z}^s$.*

Proof. Defining $\boldsymbol{z}^c$ as the class variable (which is an intuitive interpretation since the augmentation should preserve the class), [1, Thm. 3.8] applies and concludes the proof. □

D.5 Proof of Prop. 6

Proposition 6 (Fixing collapse for $\mathcal{S}^{d-1}$)**.** *Let $\boldsymbol{f}$ be a (collapsed) optimal Simplex ETF representation (i.e., $\operatorname{rank}(\mathbf{J}_{\boldsymbol{f}}) = d_c$), then the modified encoder $\boldsymbol{f}_{\mathbf{D}} = \mathbf{D} \circ \boldsymbol{f}$ with $\operatorname{rank}(\boldsymbol{f}_{\mathbf{D}}) = d = d_c + d_s$ and $\operatorname{rank}(\mathbf{D}) = d_c$ will have a rank of d_c and remain optimal w.r.t. the InfoNCE loss without collapsing.*

Proof. We want to show that multiplication with a singular diagonal matrix enables $\boldsymbol{f}$ to be full-rank, while yielding the same representation after ℓ_2-normalization. Since $\boldsymbol{z}^s$ is collapsed, we can assume w.l.o.g. that $\boldsymbol{z}^s = \mathbf{0}_s$. Assume that $\mathbf{D} = \operatorname{diag}(\mathbf{D}^c; \mathbf{D}^s)$, where $\mathbf{D}^c = \alpha_c \mathbf{I}_{d_c}, \mathbf{D}^s = \mathbf{0}_s$ ($(\cdot)|_{\boldsymbol{f}}$ means the expression evaluated for a specific encoder $\boldsymbol{f}$)

$$\|\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{z}}_j\|^2_{\big|_{\boldsymbol{f}}} = \left\|\hat{\boldsymbol{z}}_i^c - \hat{\boldsymbol{z}}_j^c\right\|^2 \tag{23}$$

$$\|\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{z}}_j\|^2_{\big|_{\boldsymbol{f}_{\mathbf{D}}}} = \left\| \mathbf{D} \left[\frac{\boldsymbol{z}_i}{\|\mathbf{D}\boldsymbol{z}_i\|} - \frac{\boldsymbol{z}_j}{\|\mathbf{D}\boldsymbol{z}_j\|} \right] \right\|^2 \tag{24}$$

$$= \left\| \mathbf{D}^c \left[\frac{\boldsymbol{z}_i^c}{\|\mathbf{D}^c\boldsymbol{z}_i^c\|} - \frac{\boldsymbol{z}_j^c}{\|\mathbf{D}^c\boldsymbol{z}_j^c\|} \right] \right\|^2 \tag{25}$$

$$= \left\| \alpha_c \left[\frac{\boldsymbol{z}_i^c}{\|\alpha_c\boldsymbol{z}_i^c\|} - \frac{\boldsymbol{z}_j^c}{\|\alpha_c\boldsymbol{z}_j^c\|} \right] \right\|^2 \tag{26}$$

$$= \|\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{z}}_j\|^2_{\big|_{\boldsymbol{f}}}. \tag{27}$$

Thus, we can conclude that the representation is the same, though $\boldsymbol{f}$ can be full-rank due to Lemma 2. □

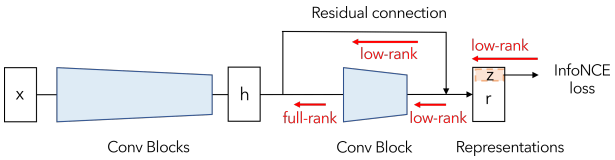

Figure 3: Setup of DirectCLR, adopted from Jing et al. [11].

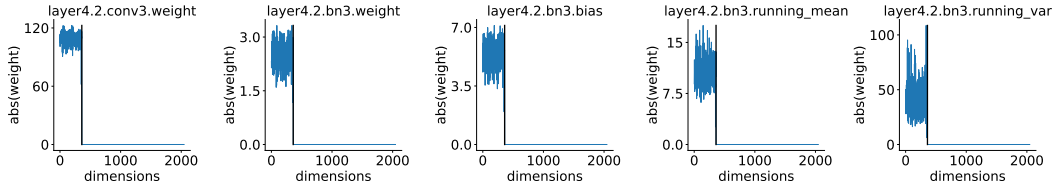

Figure 4: Absolute value of the weights in the last layer of the last Conv block of the ResNet50. The black line indicates the number of dimensions d used for training. All weights past d are driven to zero by weight decay, as this is the only explicit training signal they receive.

E Analysis of DirectCLR [11]

Jing et al. [11] also work towards understanding the role of the projection head in contrastive learning, and suggest a new method (DirectCLR) which they claim to not use an explicit trainable projector. By using DirectCLR, they avoid discarding the projection head, and do not observe the usual dimensionality collapse after the final layer. In this section, we analyze DirectCLR and show that it, in fact, does use a trainable projector — namely the last convolutional ResNet layer — and the linear readout is trained on top of a concatenation of the projected inputs and hidden representations. Furthermore, our experiments show that we can get better readout performance in the DirectCLR setting if only the hidden representations are used, as is customary in current SSL methods.

The core idea of DirectCLR is to train by applying the InfoNCE loss only to a part — denoted z — of the final layer denoted z, and then to evaluate by performing a linear readout on the full output —denoted r — of this layer (see Figure 3). The dimensionality z is set to $d = 360$, such that the InfoNCE loss is not applied to $2048 - 360 = 1688$ dimensions of r for training the network.

Seemingly, DirectCLR does not use a projection head since the authors do not need to discard z when training linear readout, and z is part of r. But to understand DirectCLR, we need to examine what r is actually composed of: The representation r is given by $r = h + \mathrm{conv}(h)$, where h is the representation in the layer before the last convolutional block $\mathrm{conv}()$ of the ResNet50 model.

After examining the trained weights of the last layer in the $\mathrm{conv}()$ block, we find that all dimensions which are not used for the InfoNCE loss are driven to zero by weight decay, see Fig. 4. The dimensionality of `layer4.2.conv3` is 2048×512, and we sum over the 512 channels; the other modules have a dimensionality of 2048. The fact that the weights corresponding to the dimensions not trained with the InfoNCE loss are zero, means that we can decompose r in two parts: the part where $\mathrm{conv}(h)$ is zero (true for 1688 dimensions), and the part where $\mathrm{conv}(h)$ is trained by the InfoNCE loss (true for 360 dimensions). Taken together, r is a concatenation of:

$$r = \mathrm{cat}\left(h[:d] + \mathrm{conv}(h), h[d:]\right). \tag{28}$$

The regular procedure for training a linear readout would be to use the representation in h, while DirectCLR can be understood as a concatenation of the representation before the projection head with the representation after the projection head. We examined training a linear readout on top of h instead of r and found an improved performance of $63.1 \pm 0.9\,\%$ compared to the $61.6 \pm 1.6\,\%$, corroborating the hypothesis that concatenating h with z works worse than using only h.

Thus, we find that the main difference between SimCLR and DirectCLR is the architecture of the projection head: while SimCLR uses a non-linear projection head composed of fully connected layers, DirectCLR uses a non-linear projection head composed of convolutions. While is not clear whether a convolutional projection head might have benefits over a fully connected one in some cases, the DirectCLR results suggest that a convolutional projection head is actually detrimental to performance

since their linear readout accuracy is significantly lower than what can be achieved with SimCLR (66.5%).

In summary, we found that

1. DirectCLR uses a non-linear convolutional projector which seems to work worse compared to the non-linear projector with fully connected layers commonly used in SimCLR.
2. DirectCLR trains linear a readout on a concatenation of the intermediate representation h and the projector outputs z, leading to worse performance than when following common practice and using h.

Acronyms
GT ground truth

CL Contrastive Learning

DGP Data Generating Process

ETF Equiangular Tight Frame

i.i.d. independent and identically distributed

MLP Multi-Layer Perceptron

ND negative definite
NSD negative semi-definite

PD positive definite

SSL Self-Supervised Learning

vMF von Mises-Fisher

Nomenclature
R^2 coefficient of determination
$\mathcal{L}_{\mathrm{CL}}$ contrastive loss function
$\mathcal{L}_{\mathrm{align}}$ alignment term in $\mathcal{L}_{\mathrm{CL}}$
$\mathcal{L}_{\mathrm{uniform}}$ uniformity term in $\mathcal{L}_{\mathrm{CL}}$
$\mathcal{S}$ hypersphere
Im image space
Ker kernel space
$\boldsymbol{f}$ encoder map $\mathcal{X} \to \mathcal{Z}$
$\boldsymbol{g}$ decoder map $\mathcal{Z} \to \mathcal{X}$
$\boldsymbol{h}$ composition of encoder and decoder, i.e., $\boldsymbol{f} \circ \boldsymbol{g}$
$\boldsymbol{\tau}$ temperature in $\mathcal{L}_{\mathrm{CL}}$

Algebra
α scalar field
$\mathbf{D}$ diagonal matrix
$\mathbf{J}$ Jacobian matrix

Latents
$\boldsymbol{\sigma}_c$ std of z^c
$\boldsymbol{\sigma}_s$ std of z^s
$\boldsymbol{z}'$ reconstructed latent vector
$\boldsymbol{z}^c$ content latent vector
$\boldsymbol{z}^s$ style latent vector
$\boldsymbol{z}$ latent vector
$\hat{\boldsymbol{z}}$ ℓ_2-normalized latent
$\mathcal{Z}^c$ content
$\mathcal{Z}^s$ style
$\mathcal{Z}$ latents
$\tilde{\boldsymbol{z}}$ positive latent vector
d_c dimensionality of $\boldsymbol{z}^c$
d_s dimensionality of $\boldsymbol{z}^s$
d dimensionality of the latent space $\mathcal{Z}$
z^c content latent single component
z^s style latent single component
z latent single component

Observations
D dimensionality of the observation space $\mathcal{X}$

M number of negative samples
$\boldsymbol{x}^-$ negative observation vector
$\boldsymbol{x}$ observation vector
$\mathcal{X}$ observation space
$\tilde{\boldsymbol{x}}$ positive observation vector

Probability theory
$\boldsymbol{\Sigma}$ covariance matrix