On the Varied Faces of Overparameterization in Supervised and Self-Supervised Learning

Matteo Gamba
KTH Royal Institute of Technology
Stockholm, Sweden

Arna Ghosh
Mila & McGill University
Montréal, QC, Canada

Kumar Krishna Agrawal
UC Berkeley
CA, USA

Blake A. Richards
Mila, Montréal Neurological Institute & McGill University
Montréal, QC, Canada
Learning in Machines and Brains Program, CIFAR
Toronto, ON, Canada

Hossein Azizpour
KTH Royal Institute of Technology
Stockholm, Sweden

Mårten Björkman
KTH Royal Institute of Technology
Stockholm, Sweden

Abstract

The quality of the representations learned by neural networks depends on several factors, including the loss function, learning algorithm, and model architecture. In this work, we use information geometric measures to assess the representation quality in a principled manner. We demonstrate that the sensitivity of learned representations to input perturbations, measured by the spectral norm of the feature Jacobian, provides valuable information about downstream generalization. On the other hand, measuring the coefficient of spectral decay observed in the eigenspectrum of feature covariance provides insights into the global representation geometry. First, we empirically establish an equivalence between these notions of representation quality and show that they are inversely correlated. Second, our analysis reveals the varying roles that overparameterization plays in improving generalization. Unlike supervised learning, we observe that increasing model width leads to higher discriminability and less smoothness in the self-supervised regime. Furthermore, we report that there is no observable double descent phenomenon in SSL with non-contrastive objectives for commonly used parameterization regimes, which opens up new opportunities for tight asymptotic analysis. Taken together, our results provide a loss-aware characterization of the different role of overparameterization in supervised and self-supervised learning.

1 Introduction

Self-supervised learning (SSL) models learn representations from large unlabeled datasets by promoting local self-consistency of the learned model function while avoiding trivial constant solutions (Zbontar et al., 2021; Chen et al., 2020; Grill et al., 2020). State-of-the-art SSL algorithms are able to learn generic features that can match the performance of supervised learning (Abbasi Koohpayegani et al., 2020). The quality of SSL representations is typically assessed via their generalization performance on downstream tasks, requiring linear probes to be trained on top of these features using labeled datasets. At present, developing a direct understanding of how the structure of SSL features

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

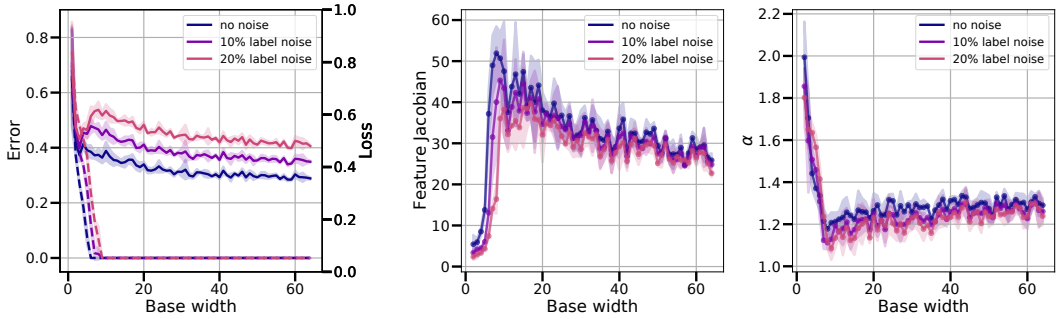

Figure 1: (Left to right) Test error (solid) and train error (dashed) for ResNet18s of increasing base width trained on CIFAR-10 as well as noisy CIFAR-10; Input Jacobian norm of features $f_{\boldsymbol{\theta}}$ learned via standard supervised learning; Spectral decay coefficient. As model size increases, the test error (left) undergoes double descent, mirrored by the feature Jacobian norm (middle). Intriguingly, the spectral decay coefficient inversely correlates with the feature Jacobian trend, decreasing until the interpolation threshold and then increasing again for overparameterized models.

affects downstream generalization is an important open problem (Agrawal et al., 2022), drawing direct parallels to understanding biological representations in the brain (Chung & Abbott, 2021).

We propose a simple information geometric framework to explore the structure of the learned representations, and their smoothness properties. In particular, for given (centered) features $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^d$, we are interested in the structure of the sample covariance $\Sigma_n = \frac{1}{n}\sum_n f_{\boldsymbol{\theta}}(\mathbf{x}_n) f_{\boldsymbol{\theta}}(\mathbf{x}_n)^T$. Recent works (Agrawal et al., 2022; He & Ozay, 2022), demonstrate that the eigenspectrum of Σ_n, $\Lambda(\Sigma_n) = \{\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d\}$, can be approximated by a power law distribution where $\lambda_i \propto i^{-\alpha}$ for $i = 1, \ldots, d$ and $\alpha > 0$. Intuitively, the *spectral-decay coefficient* α, offers a label-free measure of representation quality, where $\alpha \to \infty$ suggests *dimensionality collapse* or $\alpha \to 0$ suggests *whitening* (He & Ozay, 2022). Notably, α is a good proxy for measuring the extent of *heavy-tailed* nature in the eigenspectrum distribution, which in turn provides insights into how the signal and noise are distributed among the available d ambient dimensions. Although α is indicative of representation manifold smoothness in asymptotic regime (Stringer et al., 2019), it is unclear whether this correspondence holds in finite-dimensional settings.

In concurrent work, Gamba et al. (2023, 2022b,a) studied the role of overparameterization on the input sensitivity of neural networks' model functions, thereby quantifying the relationship between ambient dimensionality and smoothness of the learned feature space in supervised learning settings. In this work, we experimentally connect α to input sensitivity of the features $f_{\boldsymbol{\theta}}$, by studying the expected spectral norm of the input Jacobian $J = \frac{1}{n}\sum_{n=1}^{N} \|\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}\|_2$, on the training dataset $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$. In doing so, we aim to bridge the gap between notions of representation geometry and more theoretically grounded measures of representation smoothness.

Recent advances in understanding generalization of models trained with supervised learning show that deep models exhibit *double descent* under overparameterization (Belkin et al., 2018). As model size increase a neural network increases, the test error follows the classical bias-variance U-shaped curve (Geman et al., 1992) until a model is large enough to interpolate the training set. Then, as model size increases further, the test error improves again, providing a model for the remarkable generalization ability of overparameterized models (Belkin et al., 2019; Geiger et al., 2019).

Contributions We empirically explore how the geometry of learned features, under a power law assumption on their covariance eigenspectrum, relates to generalization via sensitivity of the learned representation to input perturbations. We establish a strong inverse correlation between α and input sensitivity on the training data of the underlying network function $f_{\boldsymbol{\theta}}$, and connect α to generalization performance. Our experimental contribution is two-fold: (1) in the supervised learning setting, we report that overparameterization controls α non-monotonically and that spectral decay inversely correlates with sensitivity of model function. (2) In the SSL setting, where the eigenspectrum power law was first observed, we show that the eigenspectrum of overparameterized features is characterized by lower spectral decay – corresponding to heavier eigenspectrum tails.

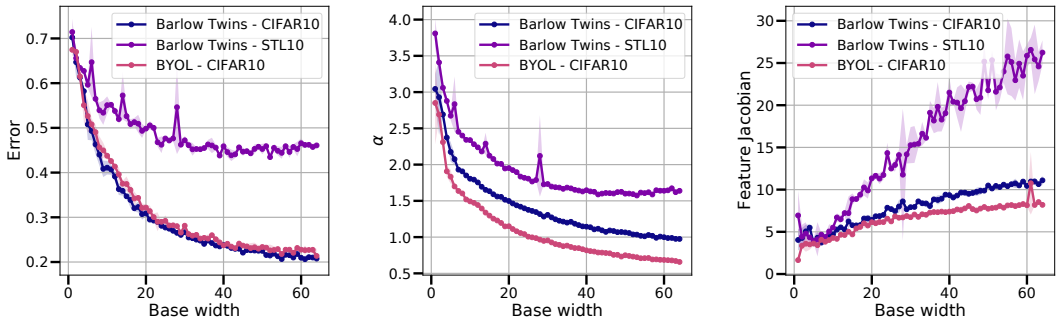

Figure 2: (Left to right) Test error for linear probes trained on SSL features with ResNet18 encoder; Input Jacobian norm of SSL features f_{θ}; Spectral decay coefficient.

2 Related work

Multi-View Self-Supervised Learning has emerged as a promising approach to learning meaningful feature representations from unlabeled data Chen et al. (2020); Zbontar et al. (2021). A key aspect of these algorithms is to leverage multiple views of the same data sample by encouraging the network to learn invariant and discriminative features that can be utilized for downstream tasks. Such algorithms have broadly been studied under the umbrella of contrastive and non-contrastive learning. Notably, non-contrastive learning algorithms, only require positively related samples. For instance, Barlow Twins (Zbontar et al., 2021), aims to enforce orthogonality among the learned features (to avoid collapse) in addition to learning to map similar images to nearby points in feature space. On the other hand, BYOL (Grill et al., 2020) employs a bootstrapped teacher-student approach. Here an online network is trained to predict the target network's output given the same input, with the target network being an exponential moving average of the online network's weights.

Smooth Interpolation has been thought to underlie generalization performance in deep neural networks. Specifically, large overparameterized models are known to learn smooth representations, thereby impacting their consistency of predictions across small input perturbations and their ability to generalize to unseen data (Gamba et al., 2023, 2022b; Novak et al., 2018). Smooth interpolation has also been shown to reconcile the classical bias-variance tradeoff understanding of double descent (Belkin et al., 2018). Furthermore, Arora et al. (2019) demonstrated that overparameterized neural networks can learn to fit noiseless data with a small number of gradient descent steps and achieve good generalization performance, implying that smoothness in the learned features plays a crucial role in the generalization capabilities of these models. Additionally, techniques like data augmentation Cubuk et al. (2018) and regularization methods such as dropout Srivastava et al. (2014) have been employed to encourage smoothness in the learned feature space, leading to improved generalization.

3 Experiments

Across our experiments, we seek to explore the role of overparameterization as we keep the optimizer fixed, but change the learning objective (i.e. supervised vs self-supervised objective). We compare the learned representations obtained with end-to-end supervised training, to SSL representations. As our backbones, we use convolutional networks with residual connections (ResNet18), which can be easily trained with both pipelines, following common practice. To control model size, following Nakkiran et al. (2019), we scale model capacity by varying the base-width of each residual block $\{1, ..., 64\}$, with experiments on CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009) and STL-10 (Coates et al., 2011). For learning objectives, we consider supervised learning as the baseline, and compare learning dynamics to Barlow Twins (Zbontar et al., 2021), BYOL (Grill et al., 2020) with 2 augmentations. Experimental details are provided in appendix A. In line with prior work, we consider the noisy-label regime to test the sensitivity and quality of features (Neyshabur et al., 2017). To measure information geometry of the learned mappings, we approximate α and the Jacobian spectral norm, per protocols elucidated in Appendix B. We also present additional results on purely convolutional models in the Appendix C.

Supervised learning setting

	CIFAR-10			CIFAR-100		
label noise	0%	10%	20%	0%	10%	20%
ConvNet	-0.61	-0.56	-0.53	-0.53	-0.73	-0.72
ResNet18	-0.62	-0.75	-0.64	-0.26	-0.13	-0.05

SSL setting

	CIFAR-10	STL-10
Barlow Twins	-0.99	-0.89
BYOL	-0.99	-0.83

Table 1: Spearman rank correlation ρ between α and the input Jacobian norm in the supervised learning (left) and SSL (right) setting.

Self-Supervised Learning A core focus of our work is understanding the scaling behaviors on pretraining with SSL objectives in the overparameterized regime. To this effect, we train Resnet-18 backbones with varying base widths using the Barlow-Twins learning objective and track the loss on training/test sets. Simultaneously, we measure α and the feature Jacobian-spectral-norm.

In Figure 2 we strikingly observe that, unlike the supervised setting, there is no observable double descent phenomenon in the self-supervised setting for the range of parameterizations considered. In particular, increasing model width monotonically improves both train and test loss. We report similar observations on STL10, and with BYOL as the pretraining objective in the Appendix C.

Equivalence of information-geometric metrics Previous works have established that higher eigenspectrum decay coefficient is indicative of a transition from discontinuous to smooth representation manifolds in infinite-dimensional settings (Stringer et al., 2019). However, it is unclear whether higher α values are indicative of smooth representations in finite-dimensional settings. Our results, both in the supervised and SSL settings, establish a strong anti-correlation between α and the spectral norm of Jacobian (Table 1). Notably, computing α takes significantly lower compute and time, compared to estimating the Jacobian, thereby offering practical computational benefits in quantifying smoothness of neural network representations.

Overparameterization and Denoising To ground our understanding of representation quality as quantified by the information geometric metrics, in this section, we evaluate downstream generalization on classification. In particular, we take backbones pretrained with Barlow-Twins and evaluate a linear probe on these frozen features with different levels of label-noise. We compare this to corresponding supervised learning baselines trained end-to-end with 20% training labels corrupted.

Following (Stephenson et al., 2020), seeking a fine-grained analysis in the overparameterized regime, we breakdown the training accuracy into three components (i) on samples with uncorrupted labels (ii) samples with corrupted labels (iii) samples with corrupted labels, evaluated on the ground truth label. In Figure 3, we visualize this for different linear evaluations. Strikingly, we observe that even with noise levels as high as 80%, the SSL pretrained models can recover up to 60% of the corrupted labels. Importantly, the supervised baseline interpolates both clean and noisy training samples, losing any label-correction ability in the overparameterized regime (base width $\omega > 9$).

4 Conclusion

Summary Our analysis reveals the different roles of overparameterization under the influence of learning objectives. In particular, we show that in self-supervised learning, the worst-case sensitivity of the model to inputs monotonically increases with overparameterization in conjunction with the emergence of heavier tails in the feature eigenspectrum, without any observable double descent phenomenon, in the range of parameterizations considered, corresponding to commonly used backbones. In contrast, overparameterized supervised features present relatively weaker tails, resulting in stronger worst-case sensitivity regularization.

Limitations Our analysis is primarily empirical in its current scope, and as such doesn't theoretically establish the absence of double-descent in SSL pretraining. Asymptotic characterizations and theoretical foundations of this learning dynamics are an exciting direction for future research.

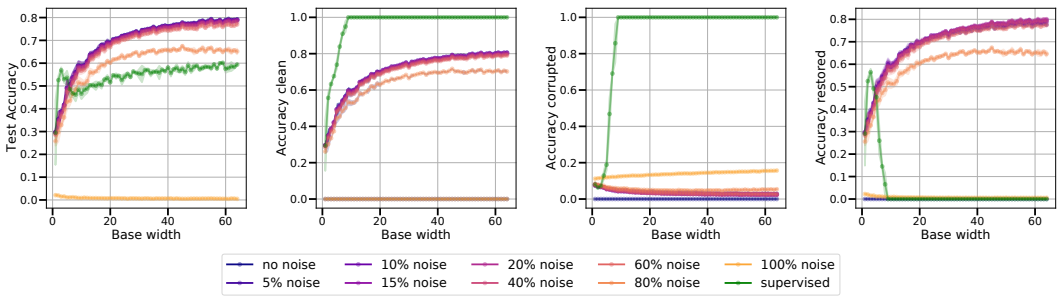

Figure 3: Downstream generalization performance on CIFAR-10 with increasing ratio of noisy labels for linear probes trained on top of frozen Barlow Twins CIFAR-10 with ResNet18 encoder features. (Left) Test accuracy for each linear probe. (Mid-left to right) We break down the training accuracy of each linear probe into accuracy on the training samples with uncorrupted labels (mid-left), and those with corrupted labels (mid-right). Finally, we assess the generality of SSL features by measuring their ability of denoise the corrupted training targets (right panel). We report for comparison analogous performance metrics for ResNet18s trained with end-to-end supervised learning on CIFAR-10 with 20% training labels randomly perturbed (green line plots).

Acknowledgments and Disclosure of Funding

The work was partially funded by Swedish Research Council project 2017-04609. Scientific computation was enabled by the supercomputing resource Berzelius provided by National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg foundation, as well as by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis partially funded by the Swedish Research Council through grant agreement no. 2022-06725. This research was generously supported by Vanier Canada Graduate scholarship (A.G.); NSERC (Discovery Grant: RGPIN-2020-05105; Discovery Accelerator Supplement: RGPAS-2020-00031; Arthur B. McDonald Fellowship: 566355-2022) and CIFAR Learning in Machines and Brains Program (B.A.R.); Canada CIFAR AI Chair program (B.A.R.). The authors also acknowledge the material support of NVIDIA in the form of computational resources.

References

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations. *Advances in Neural Information Processing Systems*, 33: 12980–12992, 2020.

Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-req: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. *Advances in Neural Information Processing Systems*, 35:17626–17638, 2022.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. doi: 10.1073/pnas.1907378117.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. Simclr: A simple framework for contrastive learning of visual representations. In *International Conference on Learning Representations*, volume 2, 2020.

SueYeon Chung and LF Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology*, 70:137–144, 2021.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2018.

Matteo Gamba, Adrian Chmielewski-Anders, Josephine Sullivan, Hossein Azizpour, and Mårten Björkman. Are all linear regions created equal? In *International Conference on Artificial Intelligence and Statistics*, pp. 6573–6590. PMLR, 2022a.

Matteo Gamba, Erik Englesson, Mårten Björkman, and Hossein Azizpour. Deep double descent via smooth interpolation. *Transactions on Machine Learning Research*, 2022b.

Matteo Gamba, Hossein Azizpour, and Mårten Björkman. On the lipschitz constant of deep networks and double descent. In *Procedings of the British Machine Vision Conference 2023*. British Machine Vision Association, 2023.

Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1):1–58, 01 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.1.1.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent - a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Bobby He and Mete Ozay. Exploring the gap between collapsed & whitened features in self-supervised learning. In *International Conference on Machine Learning*, pp. 8613–8634. PMLR, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2019.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Cory Stephenson, Abhinav Ganesh, Yue Hui, Hanlin Tang, SueYeon Chung, et al. On the geometry of generalization and memorization in deep neural networks. In *International Conference on Learning Representations*, 2020.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

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*, pp. 12310–12320. PMLR, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2018.

A Experimental setup

A.1 Supervised training

We train a family of ConvNets composed of 4 convolutional stages – each corresponding to a `[Conv, ReLU]` block followed by maxpooling with stride 2 – and 1 dense classification layer. We also train a family of ResNet18s (He et al., 2015) without batch normalization layers. Both network architectures are composed of 4 convolutional stages, in which each spatial dimension is reduced by factor of 2 and the number of learned feature maps doubles. More precisely, the convolutional stages respectively follow the progression $[\omega, 2\omega, 4\omega, 8\omega]$, where ω is the base width of the network, i.e. the number of feature maps learned at the first layer.

In our experiments, following Nakkiran et al. (2019), we vary the base width in the range $\omega = 1, \ldots, 64$. By controlling the network size through the network width, we produce a range of models presenting model-wise double descent in the test error, which captures the essence of the *benign overfitting* phenomenon(Bartlett et al., 2020) observed for large interpolating networks, while also presenting *malign overfitting* for models near the interpolation threshold. Furthermore, controlling model size through base width allows us to keep the network depth fixed, and focus our study on effective complexity of fixed-depth networks, for two network architecture families (ConvNets and ResNets).

To tune hyperparameters, we take a random validation split of size 1000 from each CIFAR training set. We train all networks with SGD with momentum 0.9, batch size 128, and fixed learning rate, set at $\eta = 5\mathrm{e}-3$ for the ConvNets and $\eta = 1\mathrm{e}-4$ for the ResNets. We train all networks for 1000 epochs. To stabilize prolonged training, we use learning rate warmup over the first 5 epochs of training, starting from a learning rate $\eta_0 = 10^{-1} \times \eta$.

A.2 SSL training

For our SSL experiments, we use a standard ResNet18 backbone with batch normalization layers as the feature encoder. Similarly to the supervised experiments, we control model size by varying the base width $\omega = 1, \ldots, 64$. Accordingly, the embedding dimensionality varies from dimension $d = 32$ to $d = 2048$. To ensure good performance, we use a non-linear projector head consisting of one hidden MLP layer with batch normalization and ReLU activations, with width matching the embedding dimensionality of the encoder.

SSL features are learned on CIFAR-10 by using Barlow Twins (Zbontar et al., 2021) and BYOL (Grill et al., 2020). At the end of SSL training, the projection layer is discarded, and the ResNet18 encoder is used to generate features $f_{\boldsymbol{\theta}}(\mathbf{x})$.

For both Barlow Twins and BYOL, SSL learning is carried out using the Adam optimizer (Kingma & Ba, 2014), with starting learning rate 0.001, and weight decay coefficients $1\mathrm{e}-5$ (Barlow Twins) and $1\mathrm{e}-6$ (BYOL), for 100 epochs (Barlow Twins) and 300 epochs (BYOL). The lambda parameter of Barlow Twins is set to $\lambda = 0.005$. The momentum parameter for BYOL is set to $\tau = 0.99$.

Once SSL features are learned, linear probes are trained for 200 epochs with SGD with base learning rate 0.1, decayed by a multiplicative factor $\gamma = 0.95$ at every epoch. The linear probes are trained without weight decay. The chosen learning rate schedule is meant to ensure that training can accommodate for fitting corrupted as well as clean labels, following Zhang et al. (2018).

A.3 Hardware specifications

Our codebase is implemented in Pytorch version 1.11, running on a local cluster equipped with NVIDIA A100 GPUs with 40GB onboard memory. Our experiments involve training 64 ConvNets and ResNets (each corresponding to a base width ω) for 1000 epochs. We use 3 random seeds, controlling network initialization and the shuffling and sampling of mini-batches from the training set. We use a dedicated random seed for generating the validation split used for hyperparameter tuning, fixed for all networks, as well as a fixed seed for corrupting the CIFAR training labels.

B Computing the feature Jacobian spectral norm

For linear operators $\mathbf{A} : (\mathbb{R}^m, \|\cdot\|_p) \to (\mathbb{R}^d, \|\cdot\|_q)$, the operator norm is defined as

$$\|A\|_{\text{op}} := \sup_{\mathbf{x}:\|\mathbf{x}\|_p \neq 0} \frac{\|A\mathbf{x}\|_q}{\|\mathbf{x}\|_p}, \tag{1}$$

where the norms $\|\cdot\|_p$ and $\|\cdot\|_q$ are respectively taken in input and feature space. Crucially, if $p = q = 2$, the operator norm can be estimated by computing the largest singular value of $\mathbf{A}$.

Computing the spectral norm Computing the spectral norm of $\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}} \in \mathbb{R}^{d \times m}$ entails two steps. First, computing the gradient $\nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}|_{\mathbf{x}=\bar{\mathbf{x}}}$ (via automatic differentiation), and then estimating its largest singular value. To perform the latter, we use a standard power method. Starting at iteration $t = 0$ with randomly initialized vectors $\tilde{\mathbf{u}}_0 \in \mathbb{R}^d$, $\tilde{\mathbf{v}}_0 \in \mathbb{R}^m$, and corresponding normalized vectors $\mathbf{u}_0 = \frac{\tilde{\mathbf{u}}_0}{\|\tilde{\mathbf{u}}_0\|_q}$, $\mathbf{v}_0 = \frac{\tilde{\mathbf{v}}_0}{\|\tilde{\mathbf{v}}_0\|_p}$, at step t we compute

$$\begin{aligned} \tilde{\mathbf{u}}_t &\leftarrow \nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}\, \mathbf{v}_{t-1} \\ \tilde{\mathbf{v}}_t &\leftarrow \mathbf{u}_t^T \nabla_{\mathbf{x}} f_{\boldsymbol{\theta}} \\ \sigma_t &\leftarrow \mathbf{u}_t^T \nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}\, \mathbf{v}_t \end{aligned} \tag{2}$$

with σ_t storing the largest singular value at convergence, defined based on a relative tolerance $1\text{e}-6$ on the size of the increments of σ_t. For large input spaces and large models $f_{\boldsymbol{\theta}}$, where instantiating the gradient might require too much memory, the power method algorithm can be efficiently computed via iterated Jacobian-vector products.

C Additional experiments

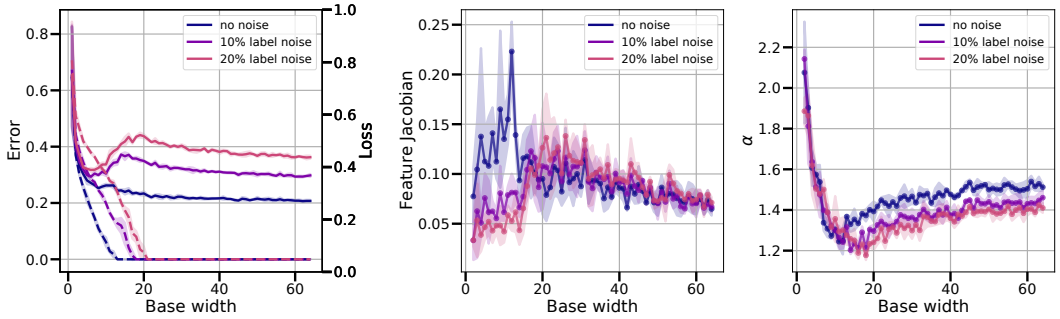

Figure 4: (Left to right) Test error (solid) and train error (dashed) for ConvNets of increasing base width trained on CIFAR-10 as well as noisy CIFAR-10; Input Jacobian norm of features $f_{\boldsymbol{\theta}}$ learned via standard supervised learning; Spectral decay coefficient. As model size increases, the test error (left) undergoes double descent, mirrored by the feature Jacobian norm (middle). Intriguingly, the spectral decay coefficient inversely correlates with the feature Jacobian trend, decreasing until the interpolation threshold and then increasing again for overparameterized models.

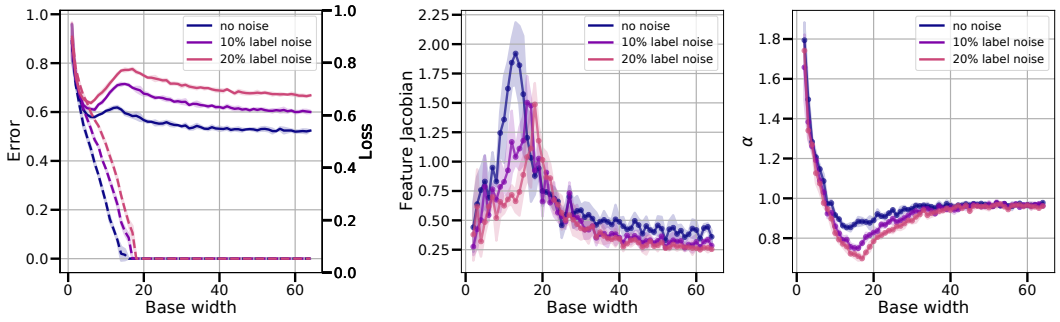

Figure 5: (Left to right) Test error (solid) and train error (dashed) for ConvNets of increasing base width trained on CIFAR-100 as well as noisy CIFAR-100; Input Jacobian norm of features f_{θ} learned via standard supervised learning; Spectral decay coefficient. As model size increases, the test error (left) undergoes double descent, mirrored by the feature Jacobian norm (middle). Intriguingly, the spectral decay coefficient inversely correlates with the feature Jacobian trend, decreasing until the interpolation threshold and then increasing again for overparameterized models.

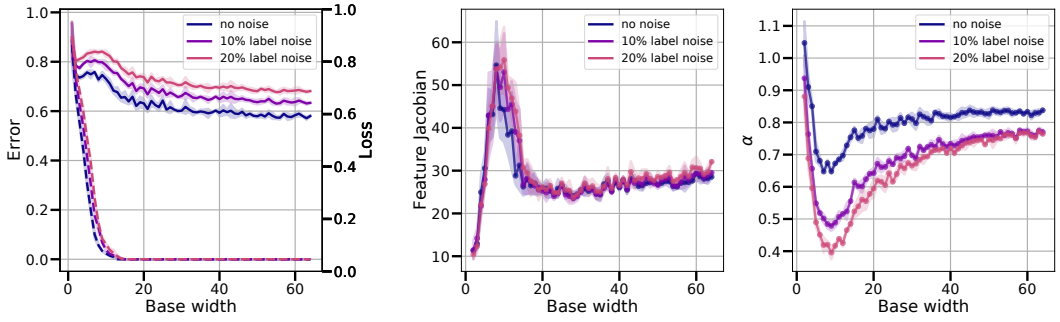

Figure 6: (Left to right) Test error (solid) and train error (dashed) for ResNet18s of increasing base width trained on CIFAR-100 as well as noisy CIFAR-100; Input Jacobian norm of features f_{θ} learned via standard supervised learning; Spectral decay coefficient. As model size increases, the test error (left) undergoes double descent, mirrored by the feature Jacobian norm (middle). Intriguingly, the spectral decay coefficient inversely correlates with the feature Jacobian trend, decreasing until the interpolation threshold and then increasing again for overparameterized models.

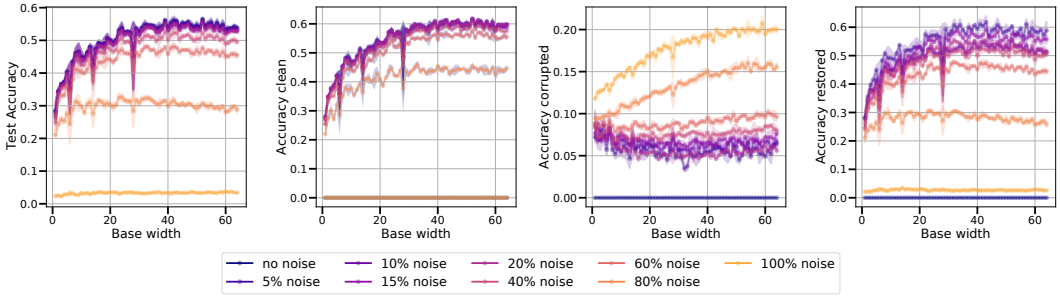

Figure 7: Downstream generalization performance on STL-10 with increasing ratio of noisy labels, for linear probes trained on top of frozen Barlow Twins STL-10 with ResNet18 encoder features. (Left) Test accuracy for each linear probe. (Mid-left to right) We break down the training accuracy of each linear probe into accuracy on the training samples with uncorrupted labels (mid-left), those with corrupted labels (mid-right). Finally, we assess the generality of SSL features by measuring their ability of denoising the corrupted training targets (right panel).

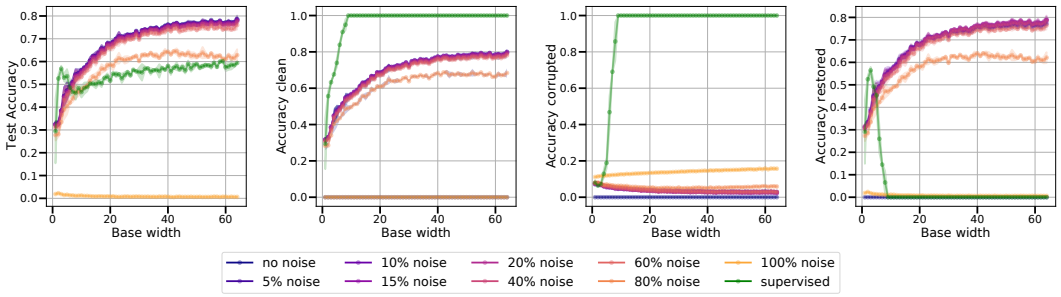

Figure 8: Downstream generalization performance on CIFAR-10 with increasing ratio of noisy labels, for linear probes trained on top of frozen BYOL CIFAR-10 with ResNet18 encoder features. (Left) Test accuracy for each linear probe. (Mid-left to right) We break down the training accuracy of each linear probe into accuracy on the training samples with uncorrupted labels (mid-left), those with corrupted labels (mid-right). Finally, we assess the generality of SSL features by measuring their ability of denoising the corrupted training targets (right panel). We report for comparison analogous performance metrics for ResNet18s trained with end-to-end supervised learning on CIFAR-10 with 20% training labels randomly perturbed (green line plots).

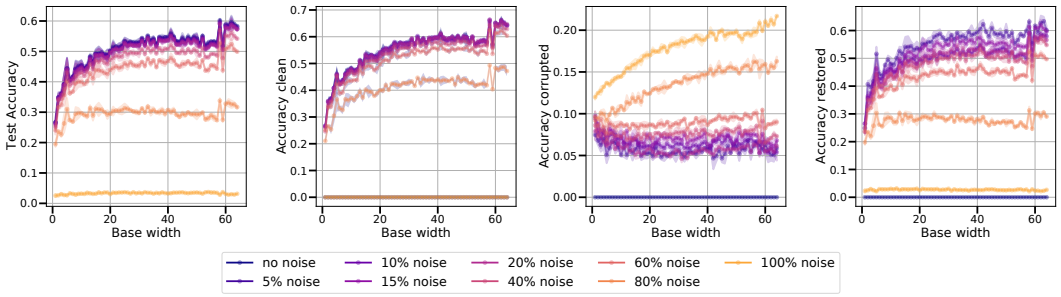

Figure 9: Downstream generalization performance on STL-10 with increasing ratio of noisy labels, for linear probes trained on top of frozen BYOL STL-10 with ResNet18 encoder features. (Left) Test accuracy for each linear probe. (Mid-left to right) We break down the training accuracy of each linear probe into accuracy on the training samples with uncorrupted labels (mid-left), those with corrupted labels (mid-right). Finally, we assess the generality of SSL features by measuring their ability of denoising the corrupted training targets (right panel).