Spectral Dynamics of Contrastive Learning with Spurious Features

Naghmeh Ghanooni GHANOONI@CS.UNI-KL.DE
RPTU, Kaiserslautern, Germany

Dennis Wagner WAGNERD@CS.UNI-KL.DE
RPTU, Kaiserslautern, Germany

Waleed Mustafa MUSTAFA@CS.UNI-KL.DE
RPTU, Kaiserslautern, Germany

Sophie Fellenz FELLENZ@CS.UNI-KL.DE
RPTU, Kaiserslautern, Germany

Anthony Widjaja Lin LIN@CS.UNI-KL.DE
RPTU, Kaiserslautern, Germany

Marius Kloft KLOFT@CS.UNI-KL.DE
RPTU, Kaiserslautern, Germany

Abstract

Contrastive learning methods are widely used to learn general-purpose representations from unlabeled data. However, they often exhibit a bias toward simple, easily learnable features—many of which may be spuriously correlated with downstream labels. This bias can limit performance, particularly for underrepresented or complex concepts. In this work, we study how such spurious correlations influence the spectral dynamics of the learned feature representations—that is, how the eigenspectrum of the feature covariance matrix evolves during training. We provide empirical and theoretical evidence that spurious features tend to dominate early spectral modes, leading to collapsed or low-rank representations that restrict downstream flexibility. To mitigate this effect, we propose a simple spectral regularization strategy that promotes high-rank representations by flattening the feature spectrum. Our method integrates seamlessly with SimCLR and improves robustness across a range of spurious correlation benchmarks. These findings highlight the importance of spectral diversity for effective self-supervised learning and suggest new directions for improving contrastive objectives.

1. Introduction

Neural networks tend to learn simple, easily detectable features in the early stages of training, a property often described as *simplicity bias* [8, 13, 20]. While such features may be predictive within the training distribution, they are often *spurious*—i.e., correlated with the labels for incidental reasons rather than semantic relevance [14, 24]. As a result, models that rely on these features may generalize poorly, especially on rare or distribution-shifted examples.

Contrastive learning (CL) has emerged as a powerful paradigm for self-supervised representation learning [5, 9, 12]. By encouraging agreement between augmented views of the same instance, CL aims to learn features that generalize across tasks. However, recent work suggests that CL models are not immune to simplicity bias and can also overfit to spurious signals present in the data [10, 32].

Because the training objective lacks knowledge of downstream labels, it may amplify correlations that are shared across augmentations but irrelevant for generalization.

In this work, we explore how spurious correlations manifest in the *spectral structure* of learned representations. Specifically, we study the eigenspectrum of the feature covariance matrix over the course of contrastive training. We show that dominant spectral directions often align with simple, spurious features that are learned early and reinforced disproportionately. This spectral imbalance results in collapsed or low-rank representations that constrain the flexibility of downstream classifiers.

To address this, we propose a spectral analysis framework and a lightweight regularization strategy that promotes a flatter spectrum. Our method increases the effective rank of representations by penalizing over-concentration in top eigenvalues, encouraging more diverse and transferable features. The regularizer is model-agnostic and integrates easily into standard contrastive pipelines.

Through a combination of theory, synthetic experiments, and real benchmarks, we analyze how spurious features shape spectral dynamics and how spectral regularization can improve robustness.

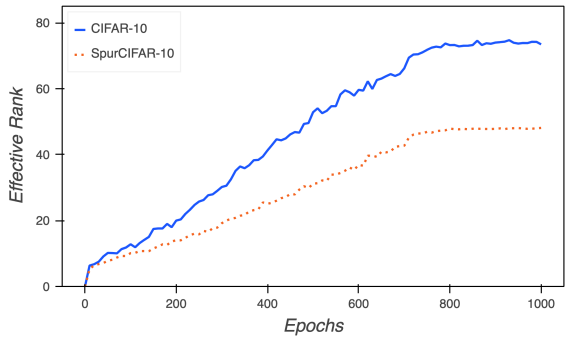
Related Work. Neural networks are known to rely on spurious features—simple patterns correlated with labels but unrelated to the task—leading to poor generalization under distribution shift [8, 23, 24, 31]. Worst-group accuracy has emerged as a key robustness metric in this context. Recent work has shown that contrastive learning and other self-supervised representation learning (SSRL) methods are also vulnerable to such spurious correlations [10, 29, 32], as they often emphasize dominant or shortcut features [5, 6, 12, 16, 30]. While some studies improve group robustness using group labels or adversarial objectives [21, 26, 27], they typically require supervision. In parallel, several works have explored the spectral properties of learned features, showing that early training emphasizes low-frequency components [13, 20, 28], and that spectral imbalance can hurt generalization [4, 11]. Although flat spectra have been linked to adversarial robustness [7], flatness alone does not guarantee transferability [1]. Our work bridges these lines by showing that reshaping the spectral dynamics of contrastive learning—toward more balanced, high-rank spectra—can mitigate spurious feature reliance without requiring group labels.

Contributions: (i) We empirically characterize how spurious correlations influence the spectral dynamics of contrastive representations and reduce effective rank, through a controlled synthetic dataset. (ii) We provide a theoretical justification for why a uniform feature spectrum minimizes generalization error over random downstream tasks. (iii) We introduce a simple spectral regularizer that improves representation diversity and downstream robustness without requiring labels or architecture changes.

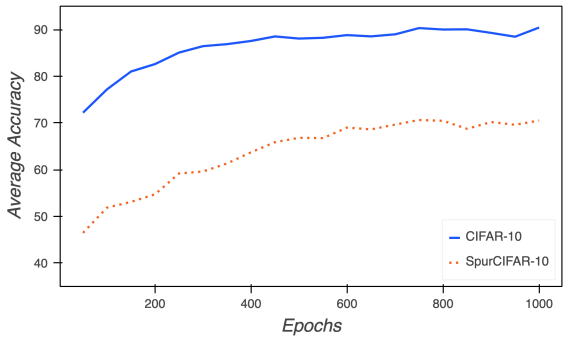
2. Spectral Imbalance Limits Downstream Flexibility

Neural networks tend to prioritize simple, easy-to-learn features during training [8, 20]. In self-supervised learning, this can lead to spurious features dominating the learned representations—especially when they are shared across augmented views but are semantically irrelevant. These features often occupy top directions in the feature covariance matrix, shaping the spectral structure in ways that reduce diversity and downstream generalization.

To understand this, we analyze the optimization dynamics of a linear predictor trained on frozen representations from a pretrained encoder, $f(\cdot)$. Given a training set $\{x_i\}_{i=1}^n$, define the corresponding feature matrix as $F = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top \in \mathbb{R}^{n \times d}$. We train a linear predictor

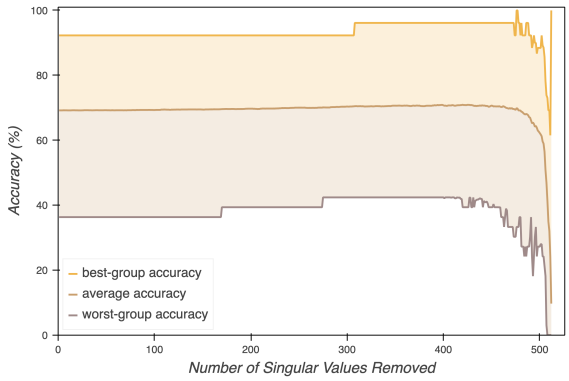


(a) Effective rank drops when training with artificial spurious features.

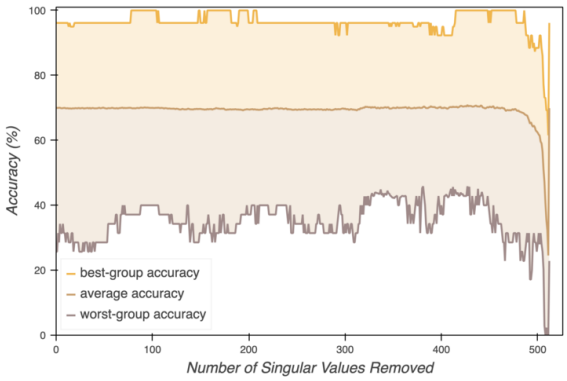


(b) Classification accuracy is lower in the presence of spurious features.

Figure 1: Spurious correlations collapse the spectrum and harm generalization. SimCLR trained on SpurCIFAR-10 learns low-rank representations (a) and suffers in classification accuracy (b), compared to CIFAR-10. This supports a link between spectral diversity and downstream robustness.



(a) Flattening after truncation steadily improves worst-group accuracy. Even flattening alone helps.



(b) Without flattening, results are inconsistent. Truncation alone does not yield robust gains.

Figure 2: Impact of spectral manipulations on SimCLR features (SpurCIFAR-10). We progressively remove low-variance components (0–512 singular values) with (left) and without (right) flattening the remaining spectrum. Flattening leads to robust gains in worst-group accuracy (from 30% to 40%) even without truncation. Specifically, to observe this gain, compare the worst-group accuracy at $x = 0$ in Figure (b) (no flattening, baseline at 30%) with $x = 0$ in Figure (a) (flattening only, 40%). This highlights that flattening alone—without truncation—improves robustness. Shaded bands indicate worst- and best-group accuracies.

$g(f(\mathbf{x})) = \langle f(\mathbf{x}), \mathbf{w} \rangle$ using gradient flow to optimize a loss function Φ, i.e. $\frac{d\mathbf{w}}{dt} = -\frac{d\Phi}{d\mathbf{w}}$. Using the chain rule of differentiation, we investigate the change of g over time:

$$\frac{dg}{dt} = \frac{dg}{d\mathbf{w}} \frac{d\mathbf{w}}{dt} = -FF^\top \frac{d\Phi}{dg} = -\sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top \frac{d\Phi}{dg},$$

where $FF^\top = \sum \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ is the eigendecomposition of the feature covariance matrix, and $\mathbf{v}_1, \cdots, \mathbf{v}_n \in \mathbb{R}^n$. Larger eigenvalues dominate learning dynamics, biasing downstream predictors toward dominant directions—regardless of whether they are semantically useful. Spurious features, being easier to learn, often dominate early and persist.

To quantify spectral imbalance, we use the *effective rank*, a principled measure based on the entropy of the singular value spectrum:

Definition 1 (Effective Rank [22]) *Let $A \in \mathbb{R}^{n \times d}$ have singular values $\sigma_1, \ldots, \sigma_Q$, with $Q = \min(n, d)$. Define:*

$$p_k = \frac{\sigma_k}{\sum_{j=1}^{Q} \sigma_j}, \quad r^{\text{eff}}(A) = \exp\left(-\sum_{k=1}^{Q} p_k \log p_k \right).$$

A high effective rank indicates a flatter spectrum and greater diversity of informative directions. A low effective rank implies that variance is concentrated in a narrow subspace—often aligned with spurious features—limiting downstream flexibility.

To illustrate this, we train SimCLR on CIFAR-10 and SpurCIFAR-10 (which introduces spurious correlations via class-specific overlays at 0.95 correlation strength). The latter produces a significantly lower effective rank and reduced test accuracy, confirming that spectral collapse accompanies overreliance on spurious cues (see Fig. 1).

In settings with spurious correlations, standard accuracy metrics can be misleading, as they are often dominated by majority or easy subgroups. To address this, we use worst-group accuracy (WG accuracy), which evaluates the model's performance on the most challenging subgroup—i.e., the group where accuracy is lowest. This metric captures robustness and fairness by revealing failure modes that are hidden by average accuracy.

We further evaluate two spectrum manipulation strategies: (1) truncating low singular values and flattening the rest, and (2) truncating without flattening. Flattening refers to reducing the disparity among the singular values of the feature matrix—i.e., making the spectrum more uniform—thereby increasing its effective rank. This encourages the representation to retain diverse informative directions rather than concentrating variance in a few dominant (and potentially spurious) components. In both strategies, we directly manipulate the feature matrix after training to achieve this uniform spectrum; importantly, this is distinct from the spectral regularizer introduced later in the paper, which encourages flattening during training. While both strategies improve worst-group (WG) accuracy, only flattening consistently enhances robustness. Even without truncation, flattening alone boosts WG accuracy from 30% to 40%. In contrast, truncation without flattening is unstable and highly sensitive to the choice of which singular values are removed (see Fig. 2).

These results support the hypothesis:

Hypothesis 2 (Spectral Diversity and Learnability) *The number of significant singular values in the feature matrix governs downstream task learnability. Higher effective rank supports more robust generalization.*

Together, these findings indicate that spectral balance is key to robust contrastive representations. Our results align with prior work on rank and robustness [1, 7, 25], while highlighting a unique trade-off: contrastive learning benefits from maintaining a well-spread spectrum that preserves both core and task-relevant signals. Flattening the spectrum encourages richer, more flexible representations—mitigating spurious bias without requiring group labels or supervision.

3. Theory: Spectral Structure and Generalization

We now formalize the connection between the spectral structure of learned representations and generalization in downstream tasks. Building on our empirical observations, we aim to understand why contrastive learning may overrepresent spurious features, and how spectral diversity can improve downstream robustness.

We consider a standard self-supervised setup where inputs $\mathbf{x} \in \mathbb{R}^{d'}$ contain both core and spurious features. An encoder $f : \mathbb{R}^{d'} \to \mathbb{R}^d$ maps inputs to d-dimensional feature representations. During contrastive training, a projection head $h(\mathbf{x}) = W_2\sigma(W_1 f(\mathbf{x}))$ is appended to the encoder, where $W_1 \in \mathbb{R}^{d_h \times d}$, $W_2 \in \mathbb{R}^{d_p \times d_h}$, and σ is a nonlinearity (e.g., ReLU). The encoder and projection head are jointly trained in an end-to-end manner using a contrastive loss, such as InfoNCE [19].

After training, the projection head is discarded, and the final learned features are stored in the matrix, $F = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top \in \mathbb{R}^{n \times d}$.

To evaluate the quality of the representations, we freeze f and train a linear classifier $g_{\mathbf{w}}(\mathbf{x}) = \langle f(\mathbf{x}), \mathbf{w} \rangle$ using labeled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $y_i \in \{-1, +1\}$, and $\mathbf{w} \in \mathbb{R}^d$ is optimized using the downstream loss, $\ell(y, g_{\mathbf{w}}(f(\mathbf{x})))$.

Recent results (e.g., Arora et al. [3]) show that the generalization error of $g_{\mathbf{w}}$, trained via gradient descent, is governed by the spectral structure of $FF^\top$. Specifically:

$$L_{\mathcal{D}}(g_{\mathbf{w}}) \leq \widetilde{O}\left(\sqrt{\frac{\mathbf{y}^\top (FF^\top)^{-1}\mathbf{y}}{n}}\right),$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. Learning is biased toward the top eigendirections of $FF^\top$; if these are aligned with spurious features, downstream robustness suffers.

Since the target task is unknown at pretraining time, we model transferability using a distribution over downstream tasks. Each task is defined by a latent direction $\mathbf{v} \in \mathbb{R}^d$, drawn from a distribution ρ. The label probabilities are given by $\mathbb{P}(Y_i = +1 \mid \mathbf{v}) = \frac{1+(F\mathbf{v})_i}{2}$ and $\mathbb{P}(Y_i = -1 \mid \mathbf{v}) = \frac{1-(F\mathbf{v})_i}{2}$.

We consider the expected loss over this task distribution

$$\mathcal{L}(F) := \mathbb{E}_{\mathbf{v},Y}\left[Y^\top (FF^\top)^{-1} Y\right].$$

Our main result shows that this expected loss is minimized when $FF^\top$ has a uniform spectrum:

Theorem 3 (Informal) *Among all feature matrices with fixed total variance (i.e., fixed trace of $FF^\top$), the expected generalization loss $\mathcal{L}(F)$ is minimized when all eigenvalues of $FF^\top$ are equal.*

Intuitively, a uniform spectrum spreads information equally across all directions, ensuring that downstream tasks—no matter which direction they rely on—can be learned efficiently. The formal statement and proof are provided in Appendix B. Appendix A also gives an illustrative example involving spurious correlations, which further motivates the benefits of a uniform spectrum.

This result complements our empirical findings and motivates spectral flattening as a regularization strategy: by promoting high-rank, balanced feature spaces, we can improve robustness to spurious signals and enable broader generalization across tasks.

4. Spectral Regularization for Robust Representation Learning

Building on our theoretical and empirical findings, we introduce a simple regularization method that promotes spectral diversity during contrastive learning. We then evaluate its performance across five benchmark datasets designed to test robustness under spurious correlations.

Regularization Strategy. Let $F \in \mathbb{R}^{n \times d}$ be the encoder output. We flatten the spectrum of $FF^\top$ to increase its effective rank by introducing a regularizer: $\mathcal{R}_{\text{spec}}(FF^\top) = \sum_{i=1}^{Q} \left(\frac{\lambda_i}{\lambda_{\max}} - 1 \right)^2$, where λ_i are the eigenvalues of $FF^\top$, $Q = \min(n, d)$, and $\lambda_{\max}$ is the largest eigenvalue. This penalizes overconcentration in top directions. The final loss becomes $\mathcal{L} = \mathcal{L}_{\text{SSRL}} + \alpha \mathcal{R}_{\text{spec}}(FF^\top)$, where $\alpha \in \mathbb{R}_+$ controls regularization strength. The regularizer is applied to each mini-batch and integrates seamlessly with SimCLR.

Evaluation Setup. We apply our method to five benchmarks spanning synthetic and real-world spurious correlations: C-MNIST [2], SpurCIFAR-10 [18], CelebA [17], MetaShift [15], and Waterbirds [23]. Models are trained using the SimCLR framework with ResNet-18 or ResNet-50 encoders. After pretraining, we freeze the encoder and train a linear classifier on top using cross-entropy loss. Group-balanced training is used for the classifier to control bias, and we report both average and worst-group accuracies.

Baselines. We compare against SimCLR and SimSiam, as well as LateTVG [10], a recent SSRL method targeting spurious robustness. All values for our method are averaged over five random seeds.

Results. Our method improves both worst-group and average accuracy across all benchmarks, outperforming LateTVG when applied to SimCLR despite using lower-dimensional (512) embeddings. It remains competitive with LateTVG on SimSiam, while requiring no architecture-specific changes. Baseline results are taken from Hamidieh et al. [10]; all our results are provided in Appendix D.

References

[1] Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. *Advances in Neural Information Processing Systems*, 36:47032–47051, 2023.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[4] N Bansal, X Chen, and Z Wang. Can we gain more from orthogonality regularizations in training deep cnns? *arXiv preprint arXiv:1810.09102*, 2018.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[7] Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems*, 35:33054–33065, 2022.

[8] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[9] 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.

[10] Kimia Hamidieh, Haoran Zhang, Swami Sankaranarayanan, and Marzyeh Ghassemi. Views can be deceiving: Improved ssl through feature space augmentation. *arXiv preprint arXiv:2406.18562*, 2024.

[11] Junlin He, Jinxiao Du, and Wei Ma. Preventing dimensional collapse in self-supervised learning via orthogonality regularization. *arXiv preprint arXiv:2411.00392*, 2024.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[13] Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

[15] Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. *arXiv preprint arXiv:2202.06523*, 2022.

[16] Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021.

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[18] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

[19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[20] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

[21] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

[22] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE, 2007.

[23] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[24] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[25] Tara N Sainath, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, and Bhuvana Ramabhadran. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6655–6659. IEEE, 2013.

[26] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173. PMLR, 2019.

[27] Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34:18225–18240, 2021.

[28] Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? on the role of simplicity bias in class collapse and feature suppression. In *International Conference on Machine Learning*, pages 38938–38970. PMLR, 2023.

[29] Haotian Ye, James Zou, and Linjun Zhang. Freeze then train: Towards provable representation learning under spurious correlations and feature noise. In *International Conference on Artificial Intelligence and Statistics*, pages 8968–8990. PMLR, 2023.

[30] 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.

[31] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[32] Weicheng Zhu, Sheng Liu, Carlos Fernandez-Granda, and Narges Razavian. Making self-supervised learning robust to spurious correlation via learning-speed aware sampling. *arXiv preprint arXiv:2311.16361*, 2023.

Appendix A. Example Motivating Importance of Uniforming the Spectrum

We present a simple example to highlight the role of the eigenspectrum of the feature matrix. Let F be a fixed feature matrix with orthonormal eigenvectors v^+, v^-, and v^s, corresponding to eigenvalues λ^+, λ^-, and λ^s, respectively. Here, v^+ and v^- represent class-discriminative directions for labels $+1$ and -1, while v^s is a spurious direction with spurious correlation strength α.

Specifically, the label generation process is as follows: define $v := \frac{1}{2}(v^+ - v^-)$. For each sample x_i, let $f_i = f(x_i)$, and define the perturbed direction:

$$v_i = \begin{cases} v, & \text{with probability } 1-\alpha, \\ \frac{1}{2}(v+v^s), & \text{with probability } \alpha. \end{cases}$$

The label $y_i \in \{\pm 1\}$ is sampled according to:

$$\mathbb{P}(y_i = +1 \mid f_i) = \frac{1 + f_i^\top v_i}{2}.$$

Let $g_i = g_w(f_i)$, and consider the squared loss:

$$\Phi(\mathbf{g}, \mathbf{y}) = \frac{1}{2}\sum_{i=1}^{n}(1 - y_i g_i)^2.$$

Lemma 4 *The expected gradient flow under the randomness of the labels satisfies:*

$$\mathbb{E}[FF^\top \cdot \nabla_{\mathbf{g}}\Phi] = \left(FF^\top \mathbf{g} - \left[\left(1-\frac{\alpha}{2}\right)\cdot\frac{1}{2}\lambda^+ v^+ - \left(1-\frac{\alpha}{2}\right)\cdot\frac{1}{2}\lambda^- v^- + \frac{\alpha}{2}\lambda^s v^s\right]\right).$$

This result shows that even a weak spurious correlation ($\alpha \ll 1$) can dominate the training dynamics if $\lambda^s \gg \lambda^+, \lambda^-$. In contrast, under a flat spectrum (i.e., uniform eigenvalues), the influence of the spurious direction scales linearly with α, making the model more robust to such noise.

Proof The first step is to compute the loss with respect to each model output g_i which is given by

$$\frac{d\Phi}{dg_i} = -2y_i(1 - y_i g_i).$$

The sources of randomness are from sampling both y and the random mixing of the spurious feature. By the law of total expectation, the expectation with respect to y and v is given by

$$\mathbb{E}_{v_i|f_i}\left[\mathbb{E}_{y_i|x_i,v_i}\left[\frac{d\Phi}{dg_i}\right]\right]$$

The inner expectation is given by

$$\mathbb{E}_{y_i|f_i,v_i}\left[\frac{d\Phi}{dg_i}\right] = -2\mathbb{E}[y_i - y_i^2 g_i] = 2g_i - 2f_i^\top v_i,$$

since $\mathbb{E}[y_i \mid f_i, v_i] = f_i^\top v_i$ and $y_i^2 = 1$. Further observe

$$\mathbb{E}_{v_i}[x_i^\top v_i] = (1-\alpha)f_i^\top v + \alpha f_i^\top\left(\frac{1}{2}(v+v^s)\right) = \left(1-\frac{\alpha}{2}\right)f_i^\top v + \frac{\alpha}{2}f_i^\top v^s.$$

Combining the above three equations and letting $\tilde{v} := \left(1 - \frac{\alpha}{2}\right) v + \frac{\alpha}{2} v^s$, we get

$$\mathbb{E}_{v_i, y_i}\left[\frac{d\Phi}{dg_i}\right] = 2g_i - 2f_i^\top \tilde{v}.$$

Stacking across all samples, let $\mathbf{g} = [g_1, \ldots, g_n]^\top$. Then:

$$\nabla_{\mathbf{g}} \Phi = (\mathbf{g} - F\tilde{v}).$$

Applying the data covariance operator $FF^\top$ gives:

$$\mathbb{E}[FF^\top \cdot \nabla_{\mathbf{g}} \Phi] = \left(FF^\top \mathbf{g} - FF^\top F\tilde{v}\right).$$

By the assumption that v^+, v^-, and v^s are orthonormal eigenvectors of $FF^\top$ with eigenvalues λ^+, λ^-, and λ^s, and $v = \frac{1}{2}(v^+ - v^-)$. Then:

$$F\tilde{v} = \left(1 - \frac{\alpha}{2}\right) \cdot \frac{1}{2} F v^+ - \left(1 - \frac{\alpha}{2}\right) \cdot \frac{1}{2} F v^- + \frac{\alpha}{2} F v^s,$$

and applying $FF^\top$:

$$FF^\top F\tilde{v} = \left(1 - \frac{\alpha}{2}\right) \cdot \frac{1}{2} \lambda^+ v^+ - \left(1 - \frac{\alpha}{2}\right) \cdot \frac{1}{2} \lambda^- v^- + \frac{\alpha}{2} \lambda_s v^s.$$

Substituting this expression concludes the proof:

$$\mathbb{E}[FF^\top \cdot \nabla_{\mathbf{g}} \Phi] = \frac{2}{n}\left(FF^\top \mathbf{g} - \left[\left(1 - \frac{\alpha}{2}\right) \cdot \frac{1}{2} \lambda^+ v^+ - \left(1 - \frac{\alpha}{2}\right) \cdot \frac{1}{2} \lambda^- v^- + \frac{\alpha}{2} \lambda^s v^s\right]\right).$$

■

Appendix B. Proof of Theorem 3

In this section, we formalize the problem setting, restate Theorem 3 and provide its proof. The dominant term, $\mathbf{y}^\top (FF^\top)^{-1} \mathbf{y}$, shows that generalization improves when the label vector $\mathbf{y}$ aligns well with the top eigenspaces of $FF^\top$. In contrastive learning, however, the downstream task is not known during pretraining, so it is unclear which directions in the feature space will ultimately be important.

To address this, we consider downstream tasks that arise by randomly sampling two latent classes $c^+, c^- \in \mathcal{C}$ according to a distribution ρ. For each such pair, we assume the existence of class-specific vectors $\mathbf{v}_{c^+}, \mathbf{v}_{c^-}$ such that the optimal linear classifier in the feature space is given by $\mathbf{v} = \mathbf{v}_{c^+} - \mathbf{v}_{c^-}$. Specifically, the class posterior is given by

$$\mathbb{P}(Y_i = +1 \mid \mathbf{v}) = \frac{1 + (F\mathbf{v})_i}{2}, \quad \mathbb{P}(Y_i = -1 \mid \mathbf{v}) = \frac{1 - (F\mathbf{v})_i}{2},$$

where $F \in \mathbb{R}^{n \times d}$ is the feature matrix.

Since downstream tasks are unknown at pretraining time, designing robust representations for contrastive learning requires optimizing for generalization over a distribution of tasks. Assuming ρ

is uniform over class pairs, we study which spectral properties of $FF^\top$ lead to improved average generalization. Specifically, we aim to minimize the expected surrogate loss:

$$\mathcal{L}(F) := \mathbb{E}_{\mathbf{v},Y}\left[Y^\top (FF^\top)^{-1} Y\right],$$

where the expectation is over random task vectors $\mathbf{v}$ and induced labels $Y \in \{\pm 1\}^n$.

The following theorem shows the optimal structure of F to enhance the generalization on a general downstream task.

Theorem 4 (Restatement of Theorem 3 (Informal)) *Let $F \in \mathbb{R}^{n\times n}$ be a feature matrix. Then, under a fixed trace constraint on $FF^\top$, the objective $\mathcal{L}(F)$ is minimized when $FF^\top$ has a uniform spectrum; that is, all eigenvalues are equal: $\lambda_1 = \lambda_2 = \cdots = \lambda_n$.*

Theorem 4 (Formal) *Let $F \in \mathbb{R}^{n\times n}$ be full-rank, and define $M := FF^\top \in \mathbb{R}^{n\times n}$. Suppose the trace is fixed, i.e., $\mathrm{Tr}(M) = \sum_{i=1}^n \lambda_i = \tau$ for some constant $\tau > 0$, where λ_i are the eigenvalues of M. Then the expected quadratic form*

$$\mathcal{L}(F) := \mathbb{E}_{Y\sim\mathcal{D}}\left[Y^\top M^{-1} Y\right]$$

is minimized when $M = \lambda I_n$, i.e., when all eigenvalues are equal.

Proof Let $M = FF^\top \in \mathbb{R}^{n\times n}$, and assume M is symmetric positive definite with eigenvalues $\lambda_1, \ldots, \lambda_n > 0$. Suppose that $Y \in \mathbb{R}^n$ is a random label vector with zero mean and isotropic covariance, i.e.,

$$\mathbb{E}[Y] = 0, \quad \mathbb{E}[YY^\top] = I_n.$$

Then the expected loss becomes:

$$\mathcal{L}(F) = \mathbb{E}_Y\left[Y^\top M^{-1} Y\right] = \mathrm{Tr}(M^{-1}).$$

This reduces the problem to minimizing $\mathrm{Tr}(M^{-1})$ subject to a fixed trace $\mathrm{Tr}(M) = \tau$, over symmetric positive definite matrices $M \in \mathbb{R}^{n\times n}$.

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of M. Since trace is invariant under orthogonal diagonalization, we have:

$$\mathcal{L}(F) = \sum_{i=1}^n \frac{1}{\lambda_i}, \quad \text{subject to} \quad \sum_{i=1}^n \lambda_i = \tau.$$

This is a classic constrained optimization problem: minimize the sum of reciprocals under a fixed sum constraint. The function $f(\lambda) = 1/\lambda$ is convex on $\lambda > 0$, and by **Jensen's inequality**, the sum $\sum_i 1/\lambda_i$ is minimized when all λ_i are equal.

Letting $\lambda_i = \tau/n$ for all i, we get the unique minimizer:

$$M = \frac{\tau}{n} I_n.$$

Thus, the spectrum is uniform and equal to τ/n.

■

Table 1: Hyperparameter settings and encoder architectures for SimCLR pretraining.

Dataset	**Encoder**	**Learning Rate**	**Batch Size**	**Weight Decay**	**Epochs**	**Regularizer α**
celebA	ResNet-50	0.01	128	1e-4	400	0.01
cmnist	ResNet-18	0.01	256	1e-4	1000	0.01
metashift	ResNet-18	0.01	256	1e-4	1000	0.005
spurcifar-10	ResNet-18	0.01	256	1e-4	1000	0.01
waterbirds	ResNet-18	0.01	256	1e-4	1000	0.01

Appendix C. Hyperparameters

We use the SimCLR framework to train ResNet encoders for all experiments. To ensure a fair comparison, we adopt the same encoder architectures as in Hamidieh et al. [10], using ResNet-18 for all datasets except CelebA, where ResNet-50 is employed. For the spectral regularization loss, we tune the regularization weight over the set $\{0.001, 0.005, 0.01, 0.05, 0.1\}$ and observe that 0.01 consistently yields the best performance. This value is therefore used in all reported experiments. A complete summary of SimCLR hyperparameters across datasets is provided in Table 1.

Appendix D. Results: Comparing to the Baselines

We report average and worst-group accuracies across five benchmark datasets in Tables 2 and 3. Our method is applied to SimCLR and compared against SimCLR, SimSiam, and their LateTVG variants. Results for baselines are taken from Hamidieh et al. [10], while our results are averaged over five random seeds. In nearly all cases, our approach outperforms SimCLR and SimCLR-LateTVG, and is competitive with SimSiam-LateTVG despite using a lower-dimensional feature space.

Table 2: Average accuracy (%) for SSRL methods (SimCLR, SimSiam), SimCLR-LateTVG, SimSiam-LateTVG, and our method, which in this experiment is applied only to SimCLR. Values are reported as mean $\pm$ standard deviation across 5 random seeds for our method. Results for other methods are taken from [10].

Dataset	SimCLR	SimSiam	SimCLR-LateTVG	SimSiam-LateTVG	Ours
C-MNIST	82.5	82.1	–	80.6	**97.0 $\pm$ 0.2**
SpurCIFAR-10	69.3	75.1	–	76.1	**80.1 $\pm$ 0.4**
CelebA	82.1	81.9	–	**88.9**	88.5 $\pm$ 0.2
MetaShift	55.1	55.8	–	70.1	**78.1 $\pm$ 0.7**
Waterbirds	47.5	50.7	–	54.8	**57.9 $\pm$ 1.3**

Table 3: Worst-group accuracy (%) for SSRL methods (SimCLR, SimSiam), SimCLR-LateTVG, SimSiam-LateTVG, and our method, which in this experiment is applied only to SimCLR. Values are reported as mean $\pm$ standard deviation across 5 random seeds for our method. Results for other methods are taken from [10].

Dataset	SimCLR	SimSiam	SimCLR-LateTVG	SimSiam-LateTVG	Ours
C-MNIST	81.7	80.7	83.8	83.1	**95.1** $\pm$ **0.3**
SpurCIFAR-10	36.5	43.4	40.4	**61.4**	59.7 $\pm$ 1.0
CelebA	76.7	77.5	82.2	83.1	**84.2** $\pm$ **0.6**
MetaShift	45.5	42.3	59.3	**79.6**	67.4 $\pm$ 0.9
Waterbirds	43.8	48.3	55.4	56.3	**56.7** $\pm$ **0.6**