Radial-VCReg: More Informative Representation Learning through Radial Gaussianization

Yilun Kuang , Yash Dagade, Deep Chakraborty, Erik Learned-Miller, Randall Balestriero, Tim G. J. Rudner, Yann LeCun

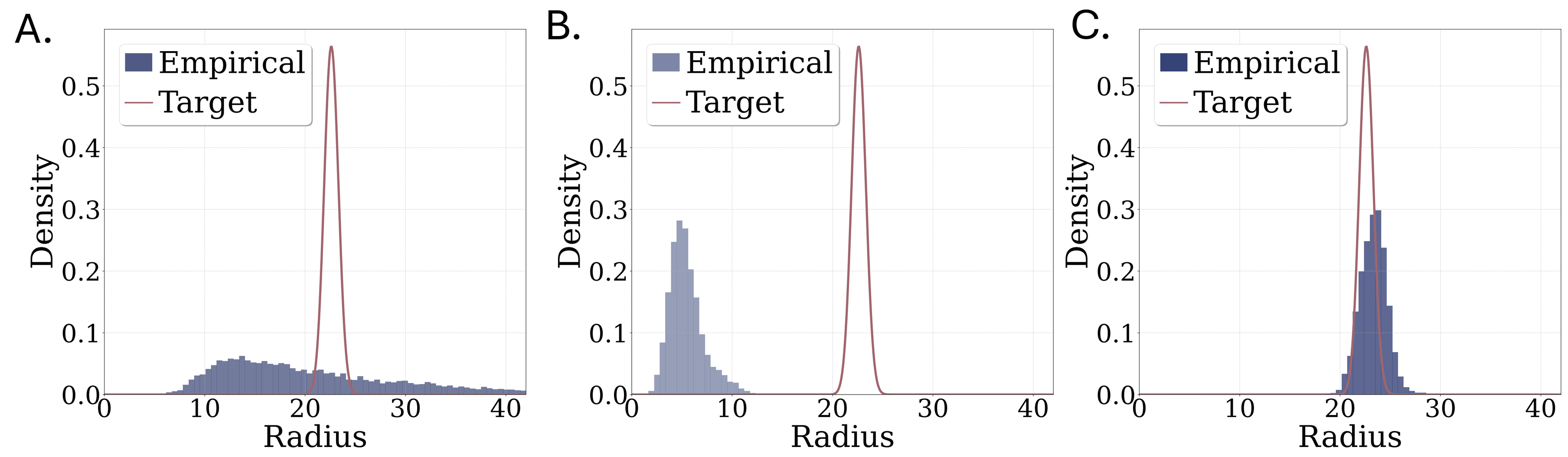
Visual Abstract

Self-supervised learning aims to learn maximally informative representations, but explicit information maximization is hindered by the curse of dimensionality. A defining property of high-dimensional Gaussian distributions is that their radial distribution follows a Chi distribution.

We propose Radial-VCReg, which augments VCReg with a Radial-Gaussianization loss. We prove that Radial-VCReg transforms a broader class of distributions toward normality compared to VCReg.

Radial-VCReg consistently improves performance by reducing higher-order dependencies and promoting more diverse and informative representations.

Fig 1.



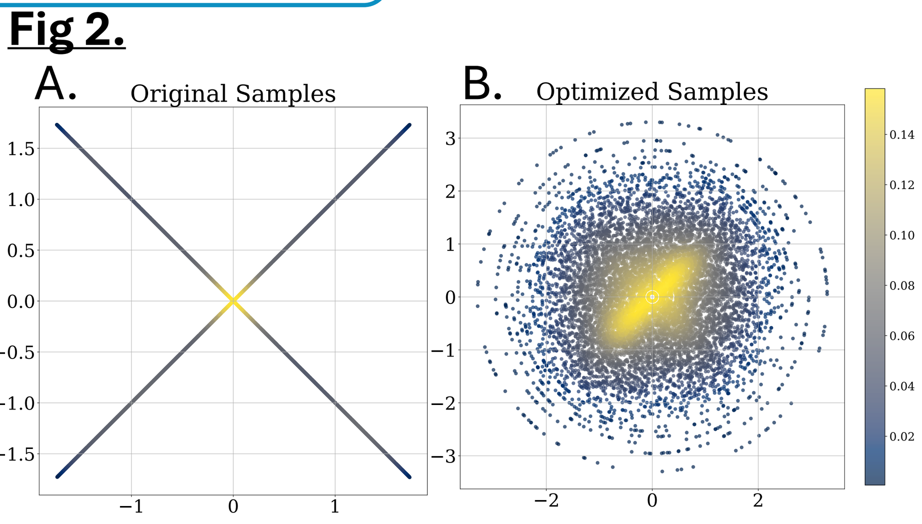
A. shows the initial feature-norm distribution at random initialization (W_1= 17.15) B. shows the representation learned with VCReg after optimization (W_1 = 8.17). C. shows the representation learned with Radial-VCReg, whose radius distribution closely matches the target Chi distribution (W_1 = 0.79).

Motivation

A d-dimensional isotropic Gaussian's radial marginal follows a Chi distribution. Enforcing this radial property together with whitening provides sufficient conditions for Gaussianity when the underlying distribution is elliptically symmetric.

Inspired by this observation, we investigate whether self-supervised representations can be made more Gaussian—and therefore more informative—by explicitly matching their radial marginal to the Chi distribution.

Fig 2.



A. The X-distribution remains unchanged under VCReg because it already has identity covariance, so VCReg cannot Gaussianize it. B. Under Radial-VICReg, the same distribution becomes spherical, elucidating that the radial loss can Gaussianize distributions that VCReg cannot transform

Radial Regularization

Radial-VICReg

For each feature z_i, define its radius $r_i = ||z_i||_2$. We want the empirical radius distribution to match the Chi distribution $\chi(d_{\text{out}})$ of an isotropic Gaussian. We approximate the KL divergence $D_{\text{KL}}\left(p_\theta(r)||p_\chi(r)\right)$ using a cross-entropy term minus an entropy term.

Where the Cross-entropy term is of the form:

$$\text{CE}(Z) = \frac{\beta_1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}||z_i||_2^2 - (d_{\text{out}} - 1)\log||z_i||_2\right).$$

And Entropy term is of the form (Vasicek m-spacing estimator)

$$H(Z) = \frac{\beta_2}{N-m}\sum_{i=1}^{N-m} log\left(\frac{N+1}{m}\left(||z_{(i+m)}||_2 - ||z_{(i)}||_2\right)\right)$$

Where we let the radii be sorted: $||z_{(1)}||_2 \le \cdots \le ||z_{(N)}||_2$.

Hence, we define radial loss as:

$$r(Z;\beta_1,\beta_2) = \text{CE}(Z) - H(Z)$$

and Radial-VICReg loss is:

$$\mathcal{L}_{\text{Radial-VICReg}}(Z,Z') = \mathcal{L}_{VICReg}(Z,Z') + r(Z;\beta_1,\beta_2) + r(Z';\beta_1,\beta_2)$$

This pushes radii toward $\chi(d_{out})$ while VICReg pushes the covariance toward identity—together moving features toward a maximum-entropy Gaussian.

Radial-VCReg Gaussianizes a Broader Class

Let $X \in R^d$ have distribution P_X with mean μ and covariance Σ. Then, the VC Reg map is

$$T_{\text{VCReg}}(x) = \Sigma^{-1/2}(x-\mu)$$

Radial VC Reg Map (Whitening + Radial CDF Transform)

Let F_r be the CDF of r and F_χ the CDF of $\chi(d)$. Then

$$T_{\text{Radial}}(x) = \frac{y}{||y||_2}F_\chi^{-1}\left(F_r(||y||_2)\right).$$

Define the sets of distributions that each map sends exactly to a standard Gaussian:

$$\mathcal{F}_{VCReg} = \{P_X: T_{\text{VCReg}}(X) \sim \mathcal{N}(0,I)\}, \mathcal{F}_{Radial} = \{P_X: T_{\text{Radial}}(X) \sim \mathcal{N}(0,I)\},$$

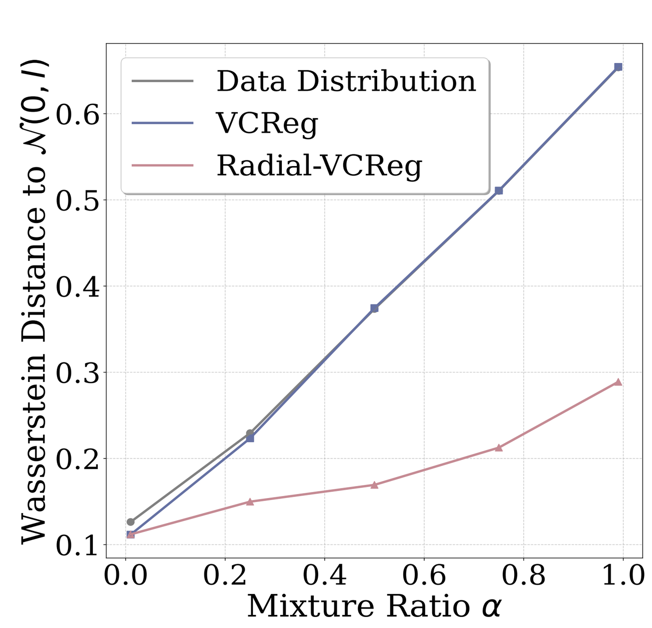
Then,

$$\mathcal{F}_{VCReg} \subsetneq \mathcal{F}_{Radial}$$

If X is elliptically symmetric, then whitening plus enforcing a Chi-distributed radius makes X exactly Gaussian. VCReg can succeed only when the radius is already Chi-distributed after whitening. Radial-VCReg explicitly corrects the radius, so it Gaussianizes strictly more distributions.

Figure 3 shows samples drawn from a mixture $\alpha X + (1-\alpha)\mathcal{N}(0,I)$, where X is the non–elliptically symmetric X-distribution that VCReg cannot Gaussianize. Radial-VICReg corrects the radial marginal and produces samples closer to $\mathcal{N}(0,I)$ across all mixture ratios, demonstrating that it Gaussianizes strictly more distributions than VCReg.

Fig 3.



Results

We assess Radial-VICReg by pretraining models with different projector dimensions on CIFAR-100 and ImageNet-10 (Tables 1–2). Radial-VICReg consistently delivers higher accuracy than VICReg at smaller projector sizes, showing gains of up to 1.5% in Top-1 accuracy across both ResNet and ViT architectures.

Table 1: **CIFAR-100 Results (Linear Probes).** The table reports the mean ± standard deviation for Top-1 and Top-5 accuracies, with the two metrics separated by a forward slash (/). All results were averaged over multiple random seeds. Hyperparameter details are provided in Appendix F.1.

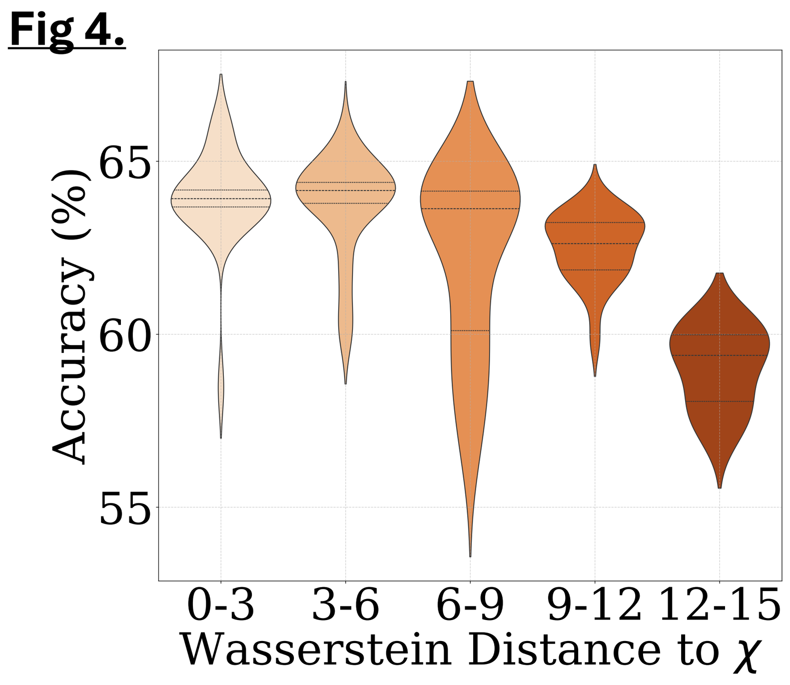
		Projector Dimension (d)	
Architecture	Method	512	2048
ResNet18	Radial-VICReg	$\mathbf{65.99 \pm 0.08}$ / $\mathbf{89.28 \pm 0.21}$	$\mathbf{68.25 \pm 0.41}$ / 90.61 ± 0.23
	VICReg	64.23 ± 0.10 / 88.32 ± 0.10	67.99 ± 0.27 / $\mathbf{90.78 \pm 0.05}$
ViT	Radial-VICReg	$\mathbf{61.33 \pm 0.29}$ / $\mathbf{87.36 \pm 0.28}$	$\mathbf{62.91 \pm 0.20}$ / $\mathbf{88.11 \pm 0.37}$
	VICReg	60.30 ± 0.21 / 86.68 ± 0.05	62.28 ± 0.33 / 87.97 ± 0.31

Table 2: **ImageNet-10 Results (Linear Probes).** The table reports the mean ± standard deviation for Top-1 and Top-5 accuracies, which are separated by a forward slash (/). All results were averaged over multiple random seeds. Hyperparameter details can be found in Appendix F.2.

Projector Dimension	512	2048	8192
Radial-VICReg	$\mathbf{94.73 \pm 0.58}$/$\mathbf{99.27 \pm 0.12}$	$\mathbf{93.93 \pm 0.31}$/99.07 ± 0.12	$\mathbf{93.33 \pm 0.70}$/$\mathbf{99.47 \pm 0.23}$
VICReg	93.20 ± 0.69/99.07 ± 0.31	93.53 ± 0.23/$\mathbf{99.47 \pm 0.31}$	$\mathbf{93.33 \pm 1.55}$/99.20 ± 0.00

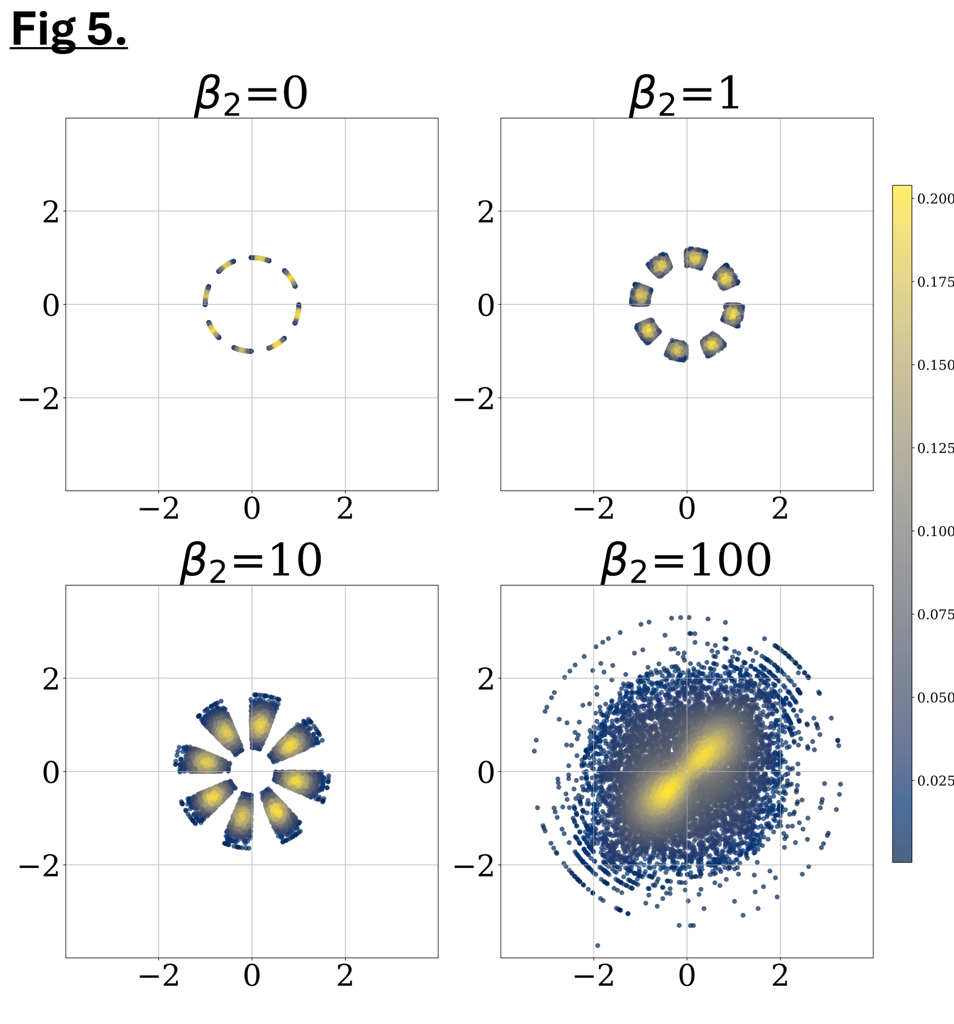
Across hyperparameter sweeps, we observe that lower Wasserstein distance to χ correlates strongly with improved downstream performance, supporting the hypothesis that Gaussianized radii yield more informative representations, as shown in Figure 4.

Fig 4.



Limitations and Future Work

Fig 5.



We highlight that some distributions cannot be Gaussianized even with Radial-VICReg. The Sunshine distribution (Fig. 6) provides such an example: despite having identity covariance and a Chi-distributed radius, varying the entropy weight β_2 cannot push it toward a true Gaussian. This figure also illustrates how changing β_2 affects the radial structure of samples. More broadly, we observe that optimal downstream performance often occurs when $\beta_1 \neq \beta_2$, even though $\beta_1 = \beta_2$ is the theoretically consistent KL estimator. Understanding this asymmetry remains open.

Because of the curse of dimensionality, perfectly Gaussian representations are infeasible to guarantee. Still, necessary but non-sufficient constraints—such as enforcing a χ-distributed radial marginal—can efficiently push features toward more isotropic, higher-entropy behavior. We view such constraints as a promising direction for future work, strengthening information-maximizing priors and enabling more expressive, disentangled representations.

Fig 6.