A. Appendix: Orthogonal Basis Functions

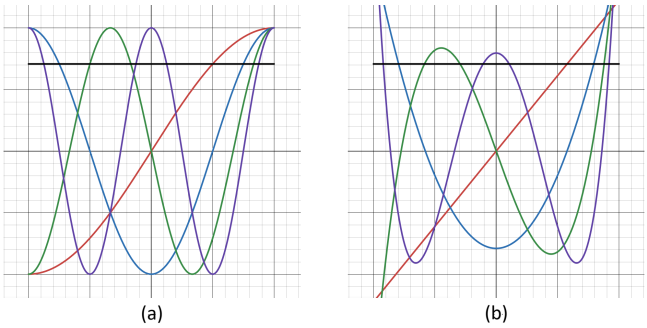

Figure 1. a. Real-valued Fourier Series b. Normalized Legendre Polynomials

DCF-Copula requires orthogonal basis functions with the following properties.

- Orthogonal over unit interval $(-1, 1)$
- Real valued, exhibiting even and odd harmonics
- Unit length L_2 norm over interval $(-1, 1)$
- All non-constant moments exhibit zero integral over $(-1, 1)$

We propose a specific normalized form of the Legendre polynomials, as well as a real-valued form of the Fourier series as seen in figure 1.

A.1. Normalized Legendre Polynomials

The Legendre Polynomials (figure 1b) are a set of real-valued orthogonal basis functions over the target interval $(-1, 1)$ with several desirable properties. Unlike the Chebyshev polynomials, the Legendre polynomials exhibit zero integral over the interval $(-1, 1)$, except trivially for the constant polynomial P_0. The following polynomials can be generated efficiently using Bonnet's recurrence. The Legendre polynomials in this form do not exhibit unit-length L_2 norm over the interval $(-1, 1)$, as such normalize these Legendre polynomials based on their L_2 as follows.

$$\phi_t(y) = \frac{P_t(y)}{||P_t||_2} \quad \text{where} \quad ||P_t||_2 = \sqrt{\int_{-1}^{1} P_t^2(y)\, dy} \tag{1}$$

A.1.1. Real-valued Fourier Series

As our sample is real-valued, one can equivalently represent the Fourier series as a sum of real-valued cos (even) and sin (odd) harmonic terms. Moreover, it is possible to simplify this to only cos terms if one makes use of the trigonometric phase identity, which is given by the following.

$$\begin{aligned} \phi_0(y) &= \frac{\sqrt{2}}{2} \\ \phi_t(y) &= cos\left(t\frac{\pi}{2}(y-1)\right) \end{aligned} \tag{2}$$

The real-valued Fourier basis functions in this form are shown in (figure 1a). These basis functions also exhibit all of our required properties.

B. Appendix: Details of Goodness of Fit

After fitting, the parametric models are evaluated using the KL-divergence between the parametric fit and the test histogram. As such, the task presented is to determine how well each of the parametric models fit to the training histogram can approximate the empirical distribution of test samples as measured by a histogram. The shaded regions present 95% confidence intervals of the layer-by-layer KL-divergence as estimated using Student's t-test. A breakdown of the steps involved with this testing procedures are as follows.

1. Compute the KL-divergence for the non-zeros samples of each filter d within the target layer of D filters, We denote this KL-divergence as KL_d. This value measures how well the trained parametric model explains the test histogram of filter d.
2. Compute the sample mean of KL-divergence values across all non-zero features in the layer.

$$\overline{\mathrm{KL}} = \frac{1}{D}\sum_{d=1}^{D} \mathrm{KL}_d \tag{3}$$

3. Estimate the standard error of the mean (SE), which quantifies the uncertainty in the estimated average KL-divergence where s is the sample standard deviation of KL-divergences,

$$s = \frac{\sigma}{\sqrt{D}}, \quad \text{where } \sigma = \sqrt{\frac{1}{D-1}\sum_{d=1}^{D}(\mathrm{KL}_d - \overline{\mathrm{KL}})^2} \tag{4}$$

4. Construct a 95% confidence interval (CI) around the mean. Since $N \geq 64$, we approximate the t-distribution with the standard normal distribution.

$$\mathrm{CI} = \overline{\mathrm{KL}} \pm z_{0.975} \cdot s, \quad \text{with } z_{0.975} \approx 1.96 \tag{5}$$

These intervals are visualized as shaded bands around the mean KL-divergence values in Figures **??** and 2. They indicate the uncertainty in the average KL-divergence for each fitted distribution and enable statistical comparison across distributions and layers. Overlapping intervals suggest no significant difference, while non-overlapping intervals indicate a statistically significant difference in goodness-of-fit.

C. Appendix: Experimental Design for Copula Inter-comparison

This appendix provides a detailed description of the methodology used in analysis of copula interdependence.

Separate Processing for Training and Testing

We have extracted training features from the training data and testing features from the testing data. First, we obtain the empirical marginal and interdependence terms strictly from the training data. Once we obtain these terms, we evaluate our model of copula interdependence by determining how well it fits the probability density of the withheld test features by using the criteria of cross entropy loss.

Probability Integral Transform

In our implementation, the empirical PIT is calculated by sorting all of the training features in the range $[0, n-1]$ in order to obtain a set of n ordered ranks. Typically, the probability integral transform converts a marginal distribution into a uniform distribution over the interval $(0, 1)$. However, in our approach, we carry out the analysis using a rescaled version of the probability integral transform that maps the feature values to the interval $(-1, 1)$. This rescaling is motivated by the fact that many standard orthogonal functions are defined on this interval, allowing us to represent the copula density in a richer and more flexible way without parametric assumptions. The modified probability integral transform is defined in the following equation.

$$F_i(x) = 2 \cdot \Pr[X_i \leq x] - 1 \tag{6}$$

Copula Density and Its Evaluation

The empirical moments $\hat{\mu}$ and copula density $\hat{c}$ are calculated in a C program that takes as input the entire training sample for the specified D features, and outputs a set of K^D empirical moments known as $\hat{\mu}$. As such the empirical moments are computed entirely from the training set. This set of moments further fully defines a model of the copula interdependence $\hat{c} : \mathbb{R}^D \rightarrow \mathbb{R}$ which is the dot product of the moments and the set of orthogonal functions.

$$\hat{c}(\vec{y}) \quad = \quad \hat{\mu} \cdot \Phi(y) \quad = \sum_{T \in \mathbb{Z}_K^D} \hat{\mu}_T \; \Phi(\vec{y}) \tag{7}$$

Now that our copula interdependence model is $\hat{c}$ is estimated from the training data, our task is to evaluate how well it models the probability density of the features from the test set. Cross entropy loss is used for the evaluation criteria.

For a given set of test features X_{test}, the transformed test features Y_{test} are calculated using the probability integral transform. Then, we use our model $\hat{c}$ to determine the predicted probability of the test features. This predicted probability is compared against the true probability of $1/N_{test}$ because empirically each test feature is equally likelihood. Therefore, the overall cross entropy evaluation is calculated using the following summation over the test set.

$$\text{Cross-Entropy} \; = \; -\frac{1}{N_{\text{test}}} \sum_{y \in Y_{test}} \log\left(\hat{c}(y)\right) \tag{8}$$

Confidence Intervals

This training and testing process is repeated 30 times for each model, dataset, and layer using a different subset of D features. For this analysis we used $D = 4$. By repeating this process 30 times, we straightforwardly calculate 95% confidence intervals using Student's t-test for the reported cross entropy loss statistics.

Archimedean copulas

Archimedean copulas make use of a Generator function $\Psi(y; \theta)$ that is invertible as follows.

$$C(y_1, y_2, \theta) \quad = \quad \Psi^{-1}\big(\; \Psi(y_1; \theta) \; + \; \Psi(y_2; \theta) \; ; \; \theta \; \big) \tag{9}$$

The Gumbel [**?**], Frank [**?**], Clayton [**?**], Ali Mikhail and Haq (AMH) [**?**], and Joe [**?**] generator functions are the most popular choices. The hyperparameter θ was determined using a formula based on either Spearman's ρ or Kendall's τ of the bivariate series. For Gumbel, Clayton and AMH, this formula is of closed form. For Joe and Frank the inverse formula is closed form. For the other methods, the inverses were calculated to high precision using binary search of the following equations.

$$\begin{aligned}
\text{Gumbel:} \quad & \theta = \frac{1}{1-\tau} \\
\text{Clayton:} \quad & \theta = \frac{2\tau}{1-\tau} \\
\text{AMH:} \quad & \theta = \frac{3\rho}{3+\rho} \\
\text{Joe:} \quad & \tau = 1 - 4\sum_{k-1}^{\infty} \frac{1}{k\big(\theta k + 2\big)\big(\theta(k-1)+2\big)} \\
\text{Frank:} \quad & \tau = 1 - \frac{4}{\theta}\left(1 - \int_0^{\theta} \frac{t}{e^t - 1} dt\right)
\end{aligned} \tag{10}$$

D. Appendix: Additional Analysis Plots

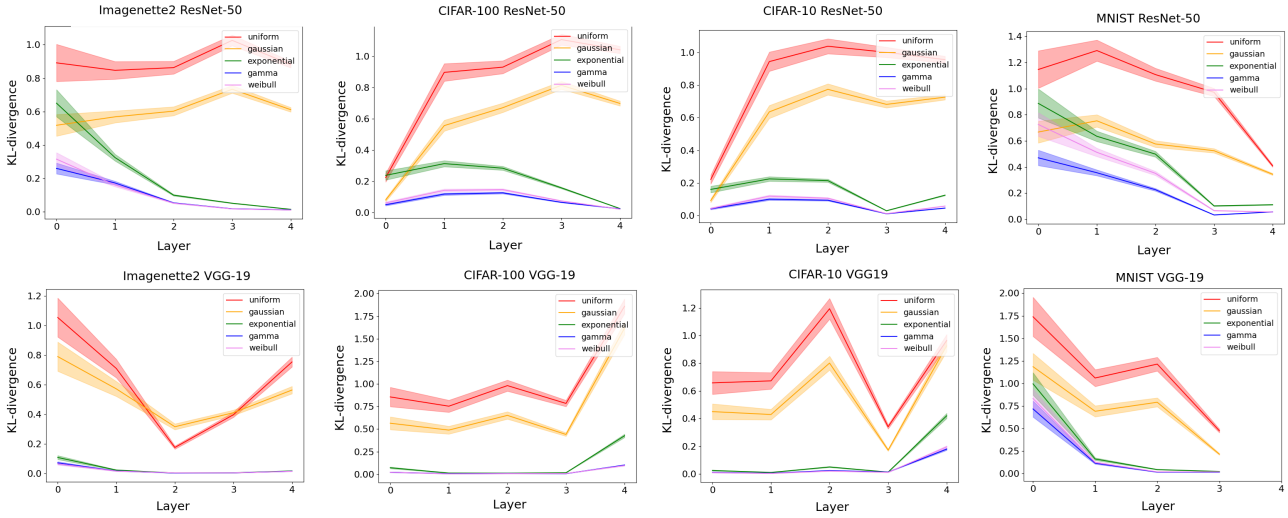

Figure 2. Overall goodness of fit of parametric marginal distributions, with shaded regions showing 2σ error bars.

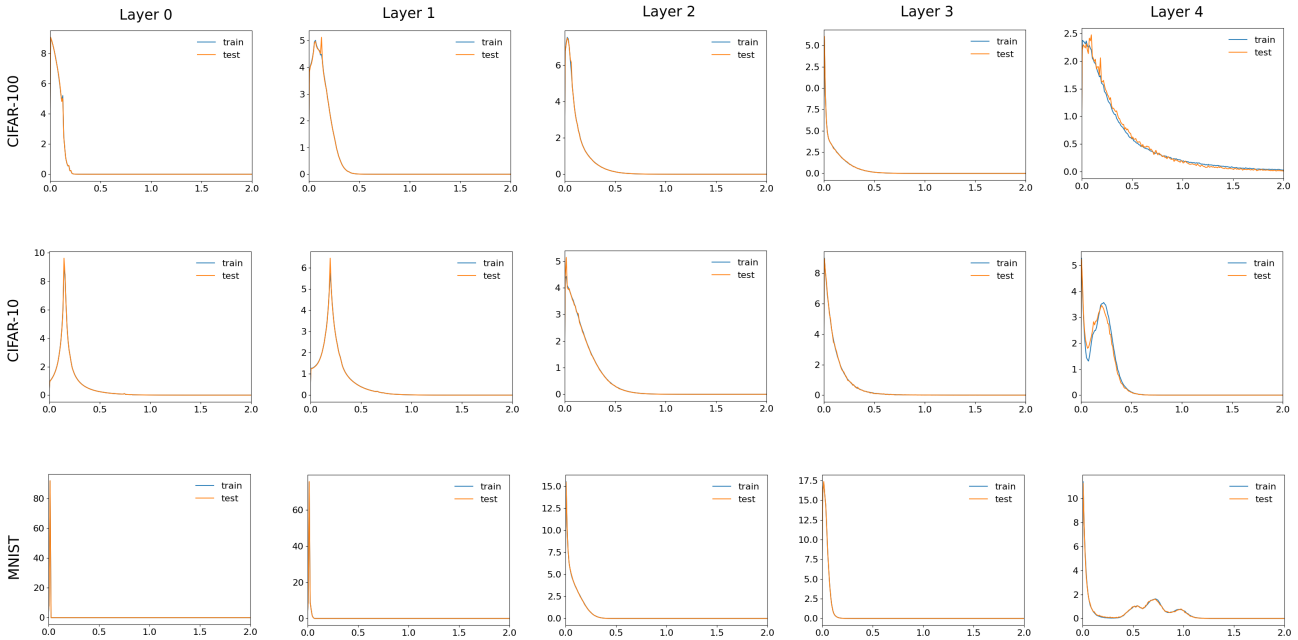

Figure 3. Supplemental examples of univariate histograms for ResNet-18 for feature 1 across CIFAR-100, CIFAR-10, and MNIST

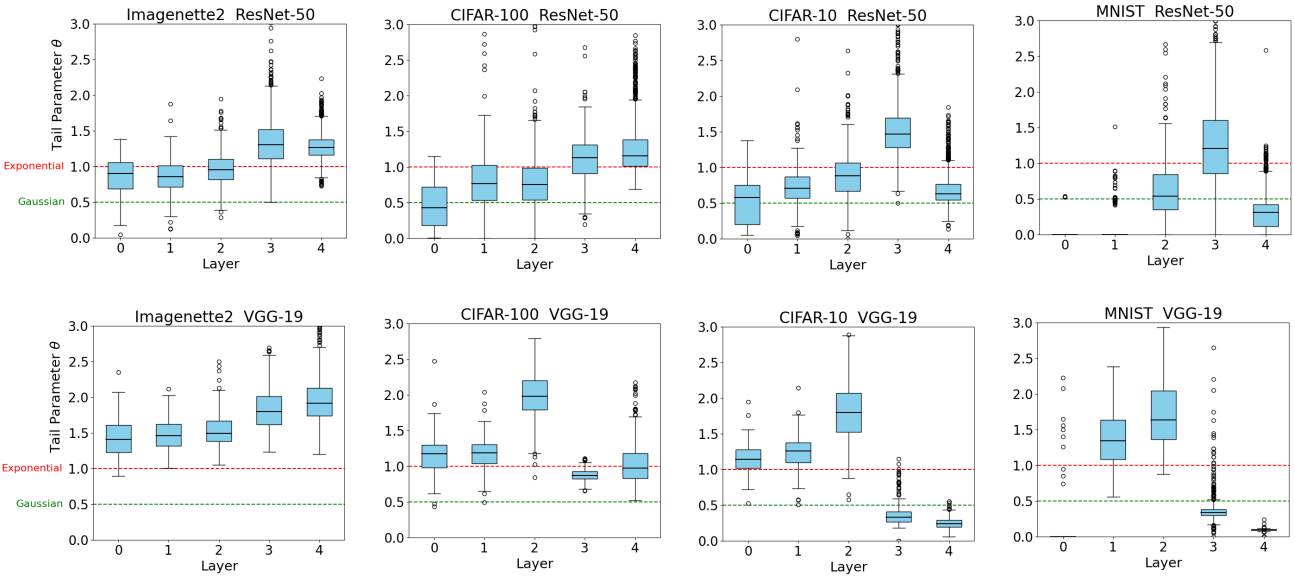

Figure 4. Estimated Weibull tail parameter θ for ResNet-50 and VGG-19 features across layers. Larger values correspond to heavier-tailed activation distributions.

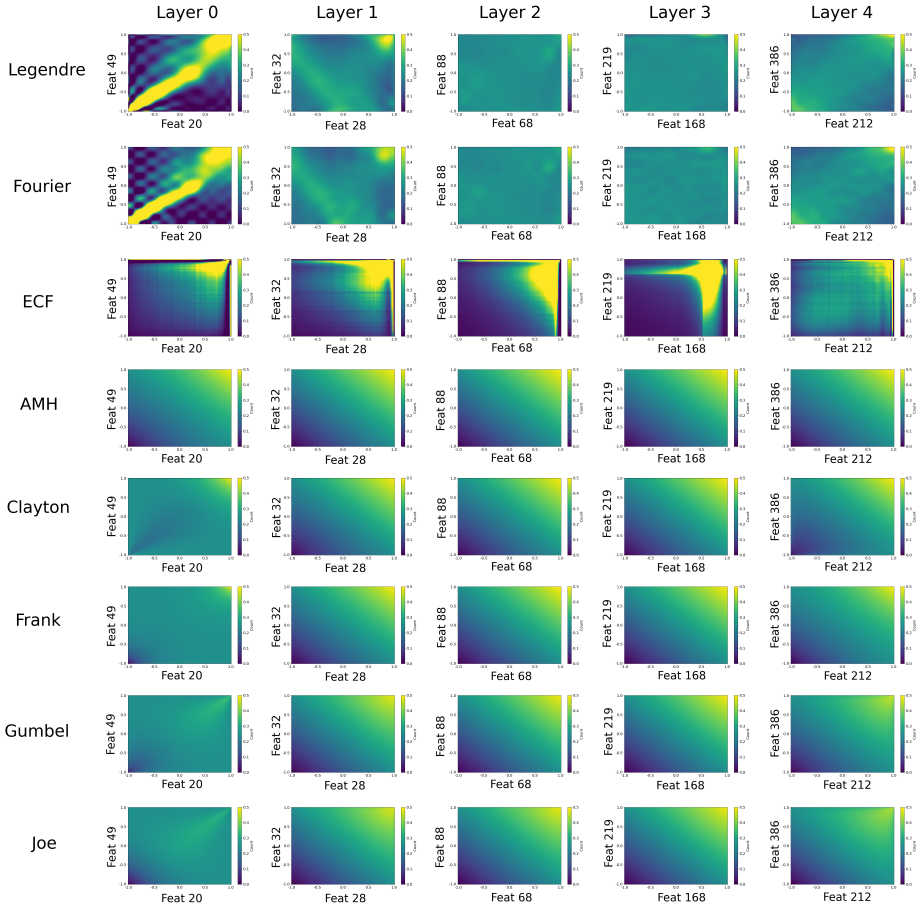

Figure 5. Comparison of copula density over random bivariate features for all methods using ResNet-18 on Imagenette2.