TIGHT FRAME CONTRACTIONS IN DEEP NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Numerical experiments demonstrate that deep neural networks classifiers progressively separate class distributions around their mean, achieving linear separability. To explain this mechanism, we introduce structured deep network architectures that can be analyzed mathematically and have high classification accuracies on complex image databases. They iterate over tight frame contractions, which apply a pointwise contraction on decomposition coefficients in a tight frame. Tight frame contractions can reduce within-class variabilities while preserving class mean separations and hence improve the Fisher discriminant ratio. Variance reduction bounds are proved for soft-thresholding contractions with Gaussian mixture models. Iterating on tight frame contractions defines deep convolutional network without bias parameters in hidden layers. We show that spatial filters do not need to be learned, and can be defined from wavelet frames. Learning frame contractions along the resulting wavelet scattering channels is sufficient to nearly reach the classification accuracies of VGG-11 and ResNet-18 on ImageNet, with no learned bias.

1 INTRODUCTION

Several numerical works (Oyallon, 2017; Papyan, 2020; Papyan et al., 2020) have shown that deep neural networks classifiers (LeCun et al., 2015) progressively concentrate each class around separated means, until the last layer where within-classes variability may nearly "collapse" (Papyan et al., 2020). The linear separability of a class mixture is characterized by the Fisher discriminant ratio (Fisher, 1936; Rao, 1948). It depends on variabilities within classes and distances between class means. The neural collapse appears through a considerable increase of the Fisher discriminant ratio during training (Papyan et al., 2020). No mathematical mechanism has yet been provided to explain this separation and concentration of probability measures.

Studying general deep networks is difficult because their architectures are complex and may vary considerably. This paper introduces a mathematical mechanism to explain this class separation, by introducing structured deep network architectures which have a high classification accuracy on complex datasets such as ImageNet. They iterate over tight frame contractions ρF, where ρ is a pointwise contraction and F is a tight frame which increases dimensionality and satisfies $F^T F = \mathrm{Id}$. The operators can reduce variances with a controlled bias on the mean, as in Stein shrinking estimations (Stein, 1956). Applying a linear classifier to a tight frame contraction defines a two-layer neural network with no bias parameter learning in the hidden layer. We call bias-free neural network any such neural network. Bias-free networks have been introduced for denoising in Mohan et al. (2019), as opposed to classification or regression. The absence of bias reduces the expressivity of such networks and we prove that they typically do not satisfy the universal approximation theorem (Pinkus, 1999; Bach, 2017). Yet their classification performances indicate that they still capture essential properties for classification, and that approximation theorems in high dimensional functional spaces may not be so central to explain deep network mathematics.

We study the properties of frame contractions depending upon the choice of contraction ρ. Whereas a soft-thresholding shrinks the amplitude of coefficients, a modulus removes their sign or their phase, and a ReLU combines both effects. For classes which are mixtures of Gaussians, we prove that if mixture centroids have sparse coefficients in the tight frame F, then a soft-thresholding considerably reduces the within-class covariance while introducing a controlled bias on the class mean. This is verified by image classification experiments with a single frame contraction. The frame contraction increases Fisher discriminant ratio by important factors.

Iterating on tight frame contractions defines bias-free deep neural networks. They alternate with orthogonal projections after each tight frame, to reduce the representation dimension. Prior information on class variabilities due to geometric groups such as translations or rotations, can be sufficient to avoid learning convolutional tight frames acting on spatial variables. A scattering transform is a cascade of wavelet convolutions with modulus contractions, which requires no learning (Bruna & Mallat, 2013). It defines a deep convolutional network with wavelet spatial filters but no linear combination of channels. Contractions along channels need to be learned. Learning 1×1 convolutional tight frame contraction across scattering channels defines a bias-free convolutional neural network. We show that it nearly reaches the classification accuracies of VGG-11 (Simonyan & Zisserman, 2015) and ResNet-18 (He et al., 2016) networks on ImageNet.

The main contributions of this paper are:

- The study of tight frame contractions to explain the reduction of within-class covariance and class mean separation in deep networks. Bounds on within-class covariance reduction are proved for Gaussian mixture models.
- The introduction of bias-free deep convolutional network architectures with cascades of tight frame contractions, which reach high accuracies on ImageNet. Learning is reduced to 1×1 convolutional tight frames, along channels created by wavelet filters.

Section 2 introduces tight frame contraction operators and two-layer networks classifiers with a single tight frame contraction. Mathematical properties are studied together with applications to image classifications on MNIST and CIFAR-10 datasets. Section 3 defines bias-free deep convolutional networks which cascade tight frame contractions. The architecture and classification accuracy of such networks are studied on ImageNet.

2 Classification with Tight Frame Contraction

2.1 Tight Frame Contractions

We briefly review properties of Fisher discriminant ratios for linear classification and introduce tight frame contractions to improve these ratios. Classification requires to identify the C components x_c of a mixture x that we shall consider to be equiprobable for simplicity. Suppose that $x \in \mathbb{R}^d$ is represented by a vector $\Phi(x)$, which is centered $\mathbb{E}(\Phi(x)) = 0$. The class averages $\mu_c = \mathbb{E}(\Phi(x_c))$ thus satisfy $\sum_c \mu_c = 0$. Let Σ_c be the covariance of $\Phi(x_c)$. One can verify that the total covariance of $\Phi(x)$ is $\Sigma_W + \Sigma_B$, where $\Sigma_W = \mathrm{Ave}_c\, \Sigma_c$ is the within-class covariance and $\Sigma_B = \mathrm{Ave}_c\, \mu_c\, \mu_c^T$ is the between-class covariance.

A linear classifier can have a small error if class means are sufficiently separated relatively to the within-class variability. The within-class covariance can be whitened and normalized to Id by transforming $\Phi(x)$ by the square root $\Sigma_W^{-\frac{1}{2}}$ of Σ_W^{-1}. All classes c, c' are highly separated if $\|\Sigma_W^{-\frac{1}{2}} \mu_c - \Sigma_W^{-\frac{1}{2}} \mu_{c'}\| \gg 1$. This separation is captured by the Fisher discriminant ratio $\Sigma_W^{-1} \Sigma_B$. We shall measure its trace:

$$C^{-1}\ \mathrm{Tr}(\Sigma_W^{-1} \Sigma_B) = \underset{c}{\mathrm{Ave}}\, \|\Sigma_W^{-\frac{1}{2}} \mu_c\|^2. \tag{1}$$

A linear classifier applies an argmax pooling on an optimized affine transformation $Wx + b$. Logistic classifiers compute W and b by minimizing a cross-entropy loss. In neural networks, this optimization is often preconditioned by an affine standardizing of $\Phi(x)$, implemented with a batch normalization (Ioffe & Szegedy, 2015). It sets the mean to zero and normalizes the variance of each coefficient.

Similar criteria based on the Fisher ratio and Linear Discriminant Analysis have been used to train deep neural networks for classification tasks as a replacement for the cross-entropy loss (Dorfer et al., 2015; Stuhlsatz et al., 2012; Sun et al., 2019; Wu et al., 2017; Sultana et al., 2018; Li et al., 2016). We however do not use eq. (1) as an optimization objective, but as a measure of classification performance, and we analyze a mechanism which allows to increase this ratio.

Tight frame contractions Optimizing $\Phi(x)$ for linear classification can be cast as a maximization of its Fisher ratio, which requires to reduce within-class covariances and preserve class separa-

tion. Shrinking estimators were introduced by Stein (Stein, 1956) to reduce the variance of mean estimators while introducing a smaller bias. Donoho & Johnstone (1994) studied shrinking by soft-thresholding in wavelet tight frames, to reduce additive noise in signals. Classification is different from noise removal because we can modify mean values as long as they remain separated between classes. In the spirit of Stein estimators, we introduce tight frame contractions to reduce within-class variance and control mean separation.

A family $\{f_m\}_{m \leq p}$ with $p \geq d$ is said to be a tight frame of $\mathbb{R}^d$ if $Fx = (\langle x, f_m \rangle)_{m \leq p}$ satisfies $F^T F = \mathrm{Id}$. A scalar non-linearity ρ is said to be contracting if $|\rho(a) - \rho(a')| \leq |a - a'|$ for all $a, a' \in \mathbb{R}$.

Definition 2.1. *A tight frame contraction in $\mathbb{R}^d$ is an operator*

$$\Phi(x) = \rho F x = \Big(\rho(\langle x, f_m \rangle)\Big)_{m \leq p} \quad \textit{with} \;\; p \geq d,$$

where $F^T F = \mathrm{Id}$ and ρ is a contraction.

A tight frame F preserves distances and since ρ is a contraction, a tight-frame contraction reduces Euclidean distances:

$$\forall x, x' \in \mathbb{R}^d \;\; , \;\; \|\Phi(x) - \Phi(x')\| \leq \|x - x'\| \; . \tag{2}$$

The contraction is only due to ρ, but its properties depend on both ρ and F. For classification, it must be optimized to reduce more the within-class covariances than the distances between each class mean $\mathbb{E}(\Phi(x_c)) = \{\mathbb{E}(\rho(\langle x, f_m \rangle))\}_m$. This is measured by the Fisher ratio (1).

Choice of contraction We study three contractions. A soft-thresholding $\rho(a) = \mathrm{sign(a)}\,(|\mathrm{a}| - \lambda, 0)$ shrinks the amplitude of a random variable a by λ in order to reduce its variances, while introducing a bias that depends on λ. It is an odd function so $\mathbb{E}(\rho(\langle x_c, f_m \rangle))$ measures the skewness of the distribution of $\langle x_c, f_m \rangle$. We shall see that a soft-thresholding can improve the Fisher ratio, but it may not separate class means that are identical. For example, if the x_c have a symmetric distribution $p(x_c) = p(-x_c)$ then $\mathbb{E}(x_c) = 0$ and $\mathbb{E}(\Phi(x_c)) = 0$. If class means are separated, a soft-thresholding can considerably reduce within-class covariance. Next section explains that it is the case if the tight frame provides sparse representations of mixture elements in each class.

To optimize the bias and variance reduction of $\langle x_c, f_m \rangle$, the threshold λ must typically be above its standard deviation, which is unknown and depends upon m. We shall make an apparently crude choice by choosing a fixed threshold λ, proportional to the standard deviation of $\langle x, f_m \rangle$ averaged over all m, which we justify with Theorem 2.2. Since $F^T F = \mathrm{Id}$ and x is standardized with $\mathbb{E}(\|x\|^2) = d$, Fx has a zero mean and $\mathbb{E}(\|Fx\|^2) = d$. We thus set $\lambda = \alpha\, d^{1/2} p^{-1/2}$, where α is of the order of 1.

A modulus $\rho(a) = |a|$ is an even contraction, which involves no parameter and has very different properties. It reduces distances by eliminating the sign. It can thus separate class means, even if they are all zero initially. This happens when the sign of $\langle x_c, f_m \rangle$ varies randomly because of a phase variability resulting from the action of a group on x_c, such as translations in images. For instance, if x_c is stationary and F is a convolutional frame with band-pass filters then $\mathbb{E}(\langle x_c, f_m \rangle) = 0$. Applying the modulus then creates non-zero means. The amplitudes $\mathbb{E}(|\langle x_c, f_m \rangle|)$ for different m and c measure the relative sparsity of the tight frame coefficients of x_c. Separating class means amounts to finding tight frames which distinguish the classes through sparsity.

A classifier applied to a ReLU output $\rho(a) = \max(a - \lambda, 0)$ can combine the properties of a soft-thresholding or a modulus in a tight frame $\{f_m\}_m$ by doubling its size to $\{f_m, -f_m\}_m$. Indeed, if $\lambda = 0$, then $|a| = \max(a, 0) + \max(-a, 0)$. For $\lambda > 0$, a soft-thresholding satisfies $\mathrm{sign(a)} \max(|\mathrm{a}| - \lambda) = \max(\mathrm{a} - \lambda, 0) - \max(-\mathrm{a} - \lambda, 0)$. In numerical applications, it typically performs as well as the best of modulus and soft-thresholding contractions, but its mathematical properties are more subtle because it can combine the properties of both.

2.2 Two-Layer Networks without Bias

We study two-layer bias-free networks that implement a linear classification on a tight frame contraction. We show that they do *not* satisfy the universal approximation theorem (Pinkus, 1999; Bach, 2017), although they reach the same image classification accuracy as standard two-layer networks.

It indicates that these functional approximation theory results may not be key mathematical properties to explain the performance of these networks. We shall see that Fisher ratio improvements rely on within-variance reductions, which can be bounded for Gaussian mixture models.

Applying a linear classifier on $\Phi(x)$ computes:

$$W\Phi(x) + b = W\rho Fx + b.$$

This two-layer neural network has no learned bias parameter in the hidden layer, and we impose that $F^T F = \mathrm{Id}$. The absence of bias parameter reduces the expressiveness compared to two-layer networks with optimized biases in the hidden layer. The following theorem proves that it does not satisfy the universal approximation theorem by proving that it cannot approximate $\|x\|^2$ with an arbitrary precision, even if the hidden layer size grows to infinity. For sake of simplicity, this theorem is proved in the case of a modulus. The extension to ReLU and soft-thresholding requires to impose an upper and lower bound on the norms of tight frame vectors, which is verified numerically. The proof of the theorem is in Appendix A.

Theorem 2.1. *If ρ is a modulus, then there exists a random process x in $\mathbb{R}^d$ which admits a probability density supported on the unit ball such that for any $p \geq d$, we have:*

$$\min_{W,F,b} \mathbb{E}\Big[\big(\|x\|^2 - W\rho Fx - b\big)^2\Big] \geq \frac{1}{180}$$

Optimization The parameters W, F and b are optimized with a stochastic gradient descent which minimizes a logistic cross-entropy loss on the output. To impose that $F^T F = \mathrm{Id}$, following the optimization of Parseval networks (Cisse et al., 2017), after each gradient update of all network parameters, we insert a second gradient step to minimize $\alpha \, \|F^T F - \mathrm{Id}\|^2$. This gradient update is:

$$F \leftarrow (1+\alpha)F - \alpha F F^T F. \tag{3}$$

To reduce the number of parameters of $W \in \mathbb{R}^{C\times p}$ we factorize $W = W' \, F^T$ with $W' \in \mathbb{R}^{C\times d}$. It amounts to reprojecting ρF in $\mathbb{R}^d$ with the semi-orthogonal frame synthesis F^T, and thus defines:

$$\Phi(x) = F^T \, \rho \, Fx.$$

A batch normalization is introduced after Φ to stabilize the learning of W'.

Image classification Image classification is first evaluated on the MNIST (LeCun et al., 2010) and CIFAR-10 (Krizhevsky, 2009) image datasets. The left part of Table 1 gives the results of logistic classifier applied to the input signal x and to $\Phi(x) = F^T \rho F x$ for 3 different contractions ρ: soft-thresholding, modulus and ReLU. The frame F is a convolution on patches of $b \times b$ with a stride of $b/2$, with $b = 14$ for MNIST and $b = 8$ for CIFAR. The tight frame F maps each patch to a vector of larger dimension, specified in Appendix C.

On each dataset, a linear classification on $\Phi(x)$ has a much smaller error than a linear classification on x, which also appears through the increases of the Fisher ratio. For MNIST, all contractions produce nearly the same classification accuracy but on CIFAR the soft-thresholding has a lower accuracy. This is due to translation variabilities. Indeed the algorithm learns a convolutional tight frame F composed mostly of band-pass filters. It implies that $\mathbb{E}(\langle x_c, f_m\rangle) \approx 0$ if x_c is nearly stationary, and as $\mathbb{E}(\Phi(\langle x_c, f_m\rangle)) \approx 0$ for a soft-thresholding non-linearity, class means are not well separated. This phenomenon is stronger on CIFAR than on MNIST, where the digit images are well centered. On the contrary, the modulus eliminates sign fluctuations and thus provide non-zero frame means in each class. The ReLU and modulus provide nearly the same results in this case.

Results obtained with ReLU and modulus are at the same level as results reported with single hidden layer architectures applied to MNIST and CIFAR, which are at 1.6% error on MNIST (Simard et al., 2003), and 25% on CIFAR-10 (Krizhevsky, 2010).

Classification of Mixtures Table 1 shows that frame concentrations $F^T \rho F x$ brings important improvements of the Fisher ratio. This improvement results from an important reduction of the within-class covariance and a controlled reduction of distances between class means. We introduce

Table 1: For MNIST and CIFAR-10, the first row gives the logistic classification error and the second row the Fisher ratio (1), for different signal representations $\Phi(x)$. Results of tight frame contractions are evaluated when ρ is a soft-thresholding (Thresh.), a modulus (Mod.) and a ReLU. A wavelet scattering transform $S(x)$ improves results, and the variability along scattering channels is reduced with $F^T\rho F$. It increases Fisher ratios and reduces errors over complex image datasets such as CIFAR.

	$\Phi(x)$	x	$F^T\rho Fx$			$S(x)$	$F^T\rho FS(x)$		
			Thresh.	Mod.	ReLU		Thresh.	Mod.	ReLU
MNIST	Error (%)	7.4	1.4	1.4	1.6	0.7	0.6	0.6	0.6
	Fisher	18.7	55.3	63.4	61.1	171.5	209.1	185.0	210.6
CIFAR	Error (%)	60.5	38.9	29.6	28.2	27.0	18.2	19.7	18.1
	Fisher	6.7	12.0	14.3	15.3	21.9	30.5	27.7	29.3

a Gaussian mixture model of each class, which explains this property for the soft-thresholding, and relates it to the sparsity of frame coefficients.

The data x has C classes x_c of mean μ_c and covariance Σ_c. The within-class covariance is $\Sigma_W = \mathrm{Ave}_c\, \Sigma_c$. After applying $\Phi(x) = F^T\rho Fx$, the mean and covariance of $\Phi(x_c)$ become $\overline{\mu}_c = \mathbb{E}(\Phi(x_c))$ and $\overline{\Sigma}_c$, and the within-class covariance becomes $\overline{\Sigma}_W = \mathrm{Ave}_c\, \overline{\Sigma}_c$. We define a Gaussian mixture model of each class x_c to evaluate the reduction of the within-class covariance.

We suppose that each x_c is a mixture of a potentially large number of Gaussian components centered in $\mu_{c,k}$ of fixed covariance $\sigma^2\mathrm{Id}$:

$$p(x_c) = \sum_k \pi_{c,k}\, \mathcal{N}(\mu_{c,k}, \sigma^2\mathrm{Id}). \tag{4}$$

This model is quite general since it amounts to covering the typical set of realizations of x_c with a union of balls centered in the $(\mu_{c,k})_k$ of radius σ. The following theorem relates the reduction of within-class covariance to the sparsity of $F\mu_{c,k}$. It relies on soft-thresholding estimation results of Donoho & Johnstone (1994). It uses a fixed threshold which uniformly reduces within-class variability along all directions.

For simplicity, we suppose that $p = d$ and hence that the tight frame is an orthogonal basis, but the result can be extended to $p \geq d$. The sparsity is expressed through the decay of sorted basis coefficients. We denote $|\langle \mu_{c,k}, f_{m_r}\rangle|$ a coefficient of rank r: $|\langle \mu_{c,k}, f_{m_r}\rangle| \geq |\langle \mu_{c,k}, f_{m_{r+1}}\rangle|$ for $1 \leq r \leq d$. The theorem imposes a condition on their amplitude decay when r increases, which is a sparsity measure. It derives upper bounds on the within-class covariance reduction and on the modifications of class means. The equivalence sign $\sim$ means that a decay is bounded by a lower and an upper constant which do not depend on the dimension d.

Theorem 2.2. *Under the mixture model hypothesis (4), we have:*

$$\mathrm{Tr}(\Sigma_W) = \mathrm{Tr}(\Sigma_M) + \sigma^2\, d, \;\; \textit{with} \;\; \mathrm{Tr}(\Sigma_M) = C^{-1} \sum_{c,k} \pi_{c,k}\, \|\mu_c - \mu_{c,k}\|^2. \tag{5}$$

If there exists $s > 1/2$ such that all ranked coefficients satisfy $|\langle \mu_{c,k}, f_{m_r}\rangle| \sim r^{-s}$, then for all $d \geq 4$ and $\lambda = \sigma\sqrt{2\log d}$, the within-class covariance of a soft-thresholding contraction satisfies:

$$Tr(\overline{\Sigma}_W) \leq 2\,\mathrm{Tr}(\Sigma_M) + O(\sigma^{1-2/s}\, \log d), \tag{6}$$

and all class means satisfy:

$$\|\mu_c - \overline{\mu}_c\|^2 = O(\sigma^{1-2/s}\, \log d). \tag{7}$$

The theorem constants do not depend upon the dimension d that is typically very large. The trace of the within-class covariance Σ_W of x is dominated by the term proportional to d. The soft-thresholding frame contraction $F^T\rho F$ reduces it to a term proportional to $\log d$, under appropriate sparsity hypotheses. If the tight frame increases the dimensionality to $p \geq d$ then the $\log d$ factor

must be replaced by $\log p$. The upper bound (7) controls the mean displacements after applying a soft-thresholding frame contraction, which is important to ensure that means remain separated. These bounds partly explains the increase of Fisher ratios, but they are not sufficient to control completely these ratios, which remains a difficult problem. For ReLU and modulus, the bound (7) on mean displacement is not valid and the separation of class means may increase strongly. In this case, characterizing the mean transformations of frame contractions raises another level of difficulties.

3 DEEP LEARNING WITH FRAME CONTRACTIONS

Next section shows that we can avoid learning the tight frame F and replace it by a wavelet frame. Cascading wavelet frame contractions defines a scattering transform. Section 3.2 shows that learning frame contractions along scattering channels considerably improves classification results.

3.1 SCATTERING BY WAVELET FRAME CONTRACTIONS

Wavelet tight frames define sparse representations of images (Mallat, 2008). Eliminating their complex phase with a modulus defines a tight frame contraction which reduces multiscale variabilities due to translations. Cascading wavelet tight frame contractions defines a scattering transform (Mallat, 2016), whose calculation is briefly reviewed. We show that it improves image classification results compared to a single frame contraction that is learned.

A wavelet frame is implemented with a filter bank which iteratively separates different frequency bands. It is computed with a low-pass filter g_0 and K complex band-pass filters g_k that are sensitive to different image orientations. If the Fourier transforms of these filters satisfy:

$$\forall \omega \in \mathbb{R}^2 \ , \ \frac{1}{2}\sum_{k=1}^{K}(|\hat{g}_k(\omega)|^2 + |\hat{g}_k(-\omega)|^2) + |\hat{g}_0(\omega)|^2 = 1,$$

then one can prove (Mallat, 2008) that $F_w x[n,k] = x \star g_k[n]$ is a complex valued tight frame representation.

Fast multiscale wavelet transforms are computed by cascading the filter bank F_w with a pooling operator with a stride of 2 (Mallat, 2008). Applying F_w with a pooling increases the scale by a factor two. To implement a frame contraction, we apply a complex modulus contraction on each band-pass channel $x \star g_k$ but not on the low-pass filter which has no complex phase. We write ρ_m this modulus contraction:

$$\rho_m F_w x[n,0] = x \star g_0[n] \ , \ \rho_m F_w x[n,k] = |x \star g_k[n]| \text{ for } k \geq 1.$$

The pooling operator P is applied after the modulus and implemented with a low-pass averaging and a subsampling with a stride of 2:

$$P\rho_m F_w x[n,0] = x \star g_0[2n] \ , \ P\rho_m F_w x[n,k] = |x \star g_k| \star g_0[2n] \text{ for } k \geq 1.$$

It reduces dimension and is orthogonal, $PP^T = \mathrm{Id}$, for an appropriate choice of low-pass filter g_0 (Mallat, 2008).

A scattering transform S of depth J and order o is a cascade of J wavelet filter banks and modulus:

$$\Phi = S = \prod_{j=1}^{J} R_j \, P \, \rho_m \, F_w.$$

Each R_j is a dimension reduction which eliminate scattering channels computed by cascading more than o modulus. We shall use a scattering transform of order $o = 2$. After the modulus contraction, we reduce dimension with $R_j P$, whereas in Section 2.2 the dimension reduction was done with F^T. However, this dimension reduction remains orthogonal: $(R_j P)(R_j P)^T = \mathrm{Id}$ for appropriate filters g_0.

Properties of scattering for image classification are studied in Bruna & Mallat (2013); Mallat (2016). If x has d pixels, then $\Phi(x)[n,k]$ is an array of images having $2^{-2J}d$ pixels at each channel k,

Table 2: Linear classification error and Fisher ratios (1) on ImageNet obtained with several representations $\Phi(x)$. The Top-5 error decays by about 20% when applying a single frame contraction along the channels of a scattering transform $S(x)$. A DeepFrame network applies a frame contraction at each layer of a wavelet scattering, which nearly reaches the accuracy of VGG-11 and ResNet-18.

	$\Phi(x)$		$S(x)$	$F^T \rho F S(x)$	DeepFrame	VGG-11	ResNet-18
ImageNet	Error (%)	Top-5	52.5	33.5	13.9	11.4	10.9
		Top-1	72.3	56.8	34.5	31.0	30.2
	Fisher		2.9	3.4	7.2	-	-

because of the J poolings P. The total number of channels k is $1+JK+J(J-1)K^2/2$. Numerical experiments are performed with wavelet filters which approximate Gabor wavelets (Bruna & Mallat, 2013), with $K = 8$ orientations at each scale. We set $J = 3$ for MNIST and CIFAR, and $J = 4$ for ImageNet, resulting in respectively 217, 651 and 1227 channels.

Table 1 gives the classification results obtained by applying a standardization across scattering channels and a logistic classifier on the scattering transform. To reduce computations, we apply on top of the standardized scattering transform an additional 3×3 orthogonal projector R, which further reduces the number of scattering channel: $RR^T = \mathrm{Id}$, followed by a spatial normalization. It improves results obtained with a single learned frame contraction by 0.7% on MNIST and by 1% on CIFAR. This improvement confirms that it is not needed to learn the spatial filters, and that wavelet frames capture the necessary discriminative spatial information.

3.2 Learning Frame Contractions Along Scattering Channels

A scattering is a cascade of spatial convolutions and modulus contractions. It involves no learning and no filtering across scattering channels. We first improve previous linear classification results on scattering by applying a single frame contraction on scattering, followed by a linear classifier. We then insert a standardized frame contraction along wavelet scattering channels at each scattering layer, which further improves classification performances. Results are also evaluated on ImageNet 2012 (Russakovsky et al., 2015).

Learning a Single Frame Contraction A scattering transform is a an array $S(x)[n, k]$ having potentially over 10^3 channels. The mean and variances of each channel are standardized with a batch normalization B_N, where affine coefficients γ and β are not learned and set respectively to 1 and 0. We apply a 3×3 convolutional orthogonal projection R to reduce the number of scattering channels. The output $RB_N S(x)[n, k]$ is an array which is also normalized at each spatial position n. A multiplicative factor sets to 1 the norm of each channel vector at each n. The tight frame contraction $F^T \rho F$ is applied to $RB_N S(x)$, which yields:

$$\Phi = \Big(F^T \rho F R\Big) B_N S.$$

Applying a linear classifier to $\Phi(x)$ defines a deep convolutional network whose learnable parameters R, F and the classifier weights W, b are optimized with stochastic gradient descent. The orthogonality properties of R and F are imposed through the gradient steps (3).

Tables 1 and 2 show that this learned tight frame contraction improves the accuracy of a wavelet scattering by 10% on CIFAR-10 and by 20% on ImageNet. The Fisher ratio (1) is similarly improved, indicating that $F^T \rho F$ reduces within-class variabilities along scattering channels.

After a scattering transform, the three contractions (soft-thresholding, modulus, ReLU) have similar performances. Indeed the wavelet scattering coefficients are locally invariant to translations, which reduces sign variabilities. Class means are already well separated by the scattering transform, which explains the performance of soft-thresholding. On ImageNet, we only evaluate the performance of soft-thresholding contractions, whose mathematical properties are simpler and partly analyzed by Theorem 2.2.

Deep Learning of Frame Contractions The previous architecture learns a single frame contraction on scattering coefficients. We now show that much higher classification performances are obtained by inserting a 1×1 frame contraction, which reduces the within-class variability of scattering channels, at each scale j. The resulting scattering frame contraction is called a DeepFrame network:

$$\Phi = \prod_{j=1}^{J} \left(F_j^T \, \rho \, F_j \, R_j \right) B_N \left(P \, \rho_m F_w \right). \tag{8}$$

It has $2J$ layers which alternate a spatial wavelet convolutional tight frame F_w with a modulus ρ_m, and a 1×1 convolutional tight frame F_j along scattering channels with a soft-thresholding ρ. In numerical experiments we set $J = 6$ for a total of 12 layers. As these tight frames increase the number of channels, dimension reductions are implemented by the pooling P which subsamples images, by a 1×1 orthogonal projection R_j which reduces the number of scattering channels, and by the frame synthesis F_j^T. If we combine F_j^T with the next wavelet convolutions F_w then we obtain separable convolutional filters along space and channels, similarly to the separable architecture in Qiu et al. (2018) computed with Fourier-Bessel filters.

This deep network does not include any learned bias parameter, except in the final linear classification layer. The learning is reduced to orthogonal 1×1 convolutional operators along scattering channels. The last layer is a logistic classifier applied to $\Phi(x)$. The network is trained by SGD and the orthogonality properties of each R_j and F_j are imposed with the gradient steps (3), leading to non-zero singular values of R_j and F_j all comprised between 0.99 and 1.01. The layer dimensions are specified in Appendix C.

Table 2 gives classification results of this DeepFrame network on ImageNet. Multiscale learning of tight frame contractions improves single frame contraction by about 20%. The resulting (Top-5, Top-1) error rates are $(13.9\%, 34.5\%)$, which nearly reaches the error rate of fully learned VGG-11 and ResNet-18 networks, respectively equal to $(11.4\%, 31.0\%)$ and $(10.9\%, 30.2\%)$[1]. This shows that cascading tight frame contractions without bias parameter yields accurate classification results on complex datasets. It does not introduce harmful simplifications and thus provides an adequate framework to study deep network properties.

4 Conclusion

Iterating on tight frame contractions defines bias-free structured deep neural networks, which can be analyzed to understand separation and concentration properties. A tight frame contraction improves linear classification by reducing within-class covariances while preserving mean separations. It appears through an improvement of the Fisher ratio. We relate reduction bounds on within-class covariance to the tight frame sparsity for soft-thresholding contractions. We showed that spatial filters do not need to be learned and can be defined as wavelets, whereas frame contractions along wavelet scattering channels brings important classification improvements. With soft-thresholding contractions, the resulting DeepFrame network has a classification accuracy which is close to VGG-11 and ResNet-18 on ImageNet.

A central mathematical issue is to characterize within-class variance reductions and mean separations across deep network layers, and prove that their Fisher discriminant ratios increase. Theorem 2.2 is not sufficient to establish this result for soft-thresholding contractions. The cases of modulus and ReLU contractions are more complex because they can strongly modify the class means, which requires appropriate hypotheses to control their separation. Although difficult, these questions on frame contractions provide a well-posed mathematical framework to study the properties of deep convolutional networks.

References

M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. Andén, E. Belilovsky, J. Bruna, V. Lostanlen, M. J. Hirn, E. Oyallon, S. Zhang, C. E. Cella,

[1]Accuracies from `https://pytorch.org/docs/stable/torchvision/models.html#classification`

and M. Eickenberg. Kymatio: Scattering transforms in python. *Journal of Machine Learning Research*, 21(60):1–6, 2020.

F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013.

M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pp. 854–863, 2017.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81 (3):425–455, 09 1994.

Matthias Dorfer, Rainer Kelz, and Gerhard Widmer. Deep linear discriminant analysis. *arXiv preprint arXiv:1511.04707*, 2015.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7): 179–188, 1936.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pp. 448–456, 2015.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Alex Krizhevsky. Convolutional deep belief networks on cifar-10, 2010.

Y. LeCun, C. Cortes, and C.J. Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Y. LeCun, Y. Bengio, and G.E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Yundong Li, Weigang Zhao, and Jiahao Pan. Deformable patterned fabric defect detection with fisher criterion-based deep learning. *IEEE Transactions on Automation Science and Engineering*, 14(2):1256–1264, 2016.

S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 3rd edition, 2008.

S. Mallat. Understanding deep convolutional networks. *Phil. Trans. of Royal Society A*, 374(2065), 2016.

Sreyas Mohan, Zahra Kadkhodaie, Eero P Simoncelli, and Carlos Fernandez-Granda. Robust and interpretable blind image denoising via bias-free convolutional neural networks. *arXiv preprint arXiv:1906.05478*, 2019.

E. Oyallon. Building a regular decision boundary with deep networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1886–1894, 2017.

V. Papyan. Traces of class/cross-class structure pervade deep learning spectra, 2020. URL `https://arxiv.org/abs/2008.11865`.

V. Papyan, X. Y. Han, and D. L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 2020.

A. Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, 1999.

Q. Qiu, X. Cheng, R. Calderbank, and G. Sapiro. DCFNet: Deep neural network with decomposed convolutional filters. *International Conference on Machine Learning*, 2018.

C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society: Series B (Methodological)*, 10(2):159–193, 1948.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

P. Simard, D. Steinkraus, and J.C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 2*, 2003.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 197–206. University of California Press, 1956.

A. Stuhlsatz, J. Lippel, and T. Zielke. Feature extraction with deep neural networks by a generalized discriminant analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 23(4):596–608, 2012.

Nazneen N Sultana, Bappaditya Mandal, and Niladri B Puhan. Deep residual network with regularised fisher framework for detection of melanoma. *IET Computer Vision*, 12(8):1096–1104, 2018.

Kai Sun, Jiangshe Zhang, Hongwei Yong, and Junmin Liu. Fpcanet: Fisher discrimination for principal component analysis network. *Knowledge-Based Systems*, 166:108–117, 2019.

Lin Wu, Chunhua Shen, and Anton Van Den Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition*, 65:238–250, 2017.