A PROOF OF PROPOSITION 2.1

We first prove the following lemma:

Lemma A.1. *If Φ is linear, then the Fisher ratio is decreased (or equal) and the optimal linear classification error is increased (or equal).*

If Φ is linear, then it is a matrix $\in \mathbb{R}^{p \times d}$. We assume that Φ has rank p (and thus $p \leq d$) for the sake of simplicity. By applying a polar decomposition on $\Phi \Sigma_W^{\frac{1}{2}}$, we can write

$$\Phi = U P \Sigma_W^{-\frac{1}{2}} ,$$

where $U \in \mathbb{R}^{p \times p}$ is symmetric positive-definite and $P \in \mathbb{R}^{p \times d}$ verifies $P P^T = \mathrm{Id}$. The within-class covariance and class means of Φx are given by

$$\begin{aligned} \overline{\Sigma}_W &= \Phi \Sigma_W \Phi^T = U^2 , \\ \overline{\mu}_c &= \Phi \mu_c = U P \Sigma_W^{-\frac{1}{2}} \mu_c . \end{aligned}$$

The Fisher ratio of Φx is thus:

$$\begin{aligned} C^{-1} \operatorname{Tr}(\overline{\Sigma}_W^{-1} \overline{\Sigma}_B) &= \operatorname*{Ave}_c \| \overline{\Sigma}_W^{-1/2} \bar{\mu}_c \|^2 \\ &= \operatorname*{Ave}_c \| P \Sigma_W^{-\frac{1}{2}} \mu_c \|^2 \\ &\leq \operatorname*{Ave}_c \| \Sigma_W^{-\frac{1}{2}} \mu_c \|^2 \\ &= C^{-1} \operatorname{Tr}(\Sigma_W^{-1} \Sigma_B), \end{aligned}$$

so Φ decreases the Fisher ratio. Besides, if (W, b) is the optimal linear classifier on Φx, then $(W\Phi, b)$ is a linear classifier on x, and thus has a larger (or equal) error than the optimal linear classifier on x.

Now, if Φ has a linear inverse Φ^{-1}, we apply the Lemma A.1 to $x' = \Phi x$ and $\Phi' = \Phi^{-1}$ (so that $\Phi' x' = x$), which concludes the proof.

Additionally, we can see from the proof of the lemma that a linear Φ preserves the Fisher ratio if and only if $\| P \Sigma_W^{-\frac{1}{2}} \mu_c \| = \| \Sigma_W^{-\frac{1}{2}} \mu_c \|$ for all c. This happens when $\Sigma_W^{-\frac{1}{2}} \mu_c$ is in the orthogonal of $\operatorname{Ker} P = \operatorname{Ker} U P = \operatorname{Ker} \Phi \Sigma_W^{\frac{1}{2}}$, which means that $\Sigma_W^{-1} \mu_c$ is in the orthogonal of $\operatorname{Ker} \Phi$. When Φ is an orthogonal projector, the orthogonal of $\operatorname{Ker} \Phi$ is the range of Φ.

B PROOF OF THEOREM 2.2

We begin by proving (3). Since $\operatorname{Tr}(\Sigma_W) = \operatorname{Ave}_c \operatorname{Tr}(\Sigma_c)$ with $\operatorname{Tr}(\Sigma_c) = \mathbb{E}(\|x_c - \mu_c\|^2)$ and x_c is a mixture of $\mathcal{N}(\mu_{c,k}, \sigma^2 \mathrm{Id})$ we get that $\operatorname{Tr}(\Sigma_c) = \operatorname{Tr}(\Sigma_M) + d\,\sigma^2$ with

$$\operatorname{Tr}(\Sigma_M) = C^{-1} \sum_k \pi_{c,k} \, \| \mu_{c,k} - \mu_c \|^2,$$

which verifies (3).

The inequalities (4) and (5) of Theorem 2.2 are derived from the following lemma which is mostly a consequence of a theorem proved by Donoho & Johnstone (1994) on soft-thresholding estimators.

Lemma B.1. *Let x be a d dimensional Gaussian vector whose distribution is $\mathcal{N}(\mu, \sigma^2 \mathrm{Id})$ with $|\mu^{(r)}| \sim r^{-s}$. For all $d \geq 4$ and $\lambda = \sigma \sqrt{2 \log d}$,*

$$\mathbb{E}(\|\rho_t(x) - \mu\|^2) = O(\sigma^{2 - 1/s} \, \log d). \tag{9}$$

Each class x_c is a mixture of several $x_{c,k}$ whose distributions are $\mathcal{N}(\mu_{c,k}, \sigma^2 \mathrm{Id})$. We first prove the theorem by applying this lemma to each $x_{c,k}$, and we shall then prove the lemma.

We apply (9) to $x = Fx_{c,k}$, $\mu = F\mu_{c,k} = \{\langle \mu_{c,k}, f_m \rangle\}_m$, and $\Phi = F^T \rho F$. Since F is orthogonal

$$\mathbb{E}(\|\Phi(x_{c,k}) - \mu_{c,k}\|^2) = \mathbb{E}(\|\rho_t F x_{c,k} - F\mu_{c,k}\|^2) = O(\sigma^{2-1/s} \log d). \tag{10}$$

Let $\overline{\mu}_c = \mathbb{E}(\Phi(x_c))$ and $\overline{\mu}_{c,k} = \mathbb{E}(\Phi(x_{c,k}))$. As we have the decomposition

$$\mathbb{E}(\|\Phi(x_{c,k}) - \mu_{c,k}\|^2) = \mathbb{E}(\|\Phi(x_{c,k}) - \overline{\mu}_{c,k}\|^2) + \|\overline{\mu}_{c,k} - \mu_{c,k}\|^2,$$

equation (10) implies that

$$\|\overline{\mu}_{c,k} - \mu_{c,k}\|^2 = O(\sigma^{2-1/s} \log d) \tag{11}$$

and

$$\mathbb{E}(\|\Phi(x_{c,k}) - \overline{\mu}_{c,k}\|^2) = O(\sigma^{2-1/s} \log d). \tag{12}$$

We first prove (5) by observing that

$$\|\mu_c - \overline{\mu}_c\|^2 = \|\sum_k \pi_{c,k}(\mu_{c,k} - \overline{\mu}_{c,k})\|^2 \leq \Big(\sum_k \pi_{c,k} \|\mu_{c,k} - \overline{\mu}_{c,k}\|\Big)^2$$

It results from (11) that

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

which proves (5).

As in the proof of (3), we verify that

$$\mathrm{Tr}(\overline{\Sigma}_W) = \mathrm{Tr}(\overline{\Sigma}_M) + C^{-1} \sum_{c,k} \pi_{c,k}\, \mathbb{E}(\|\Phi(x_{c,k}) - \overline{\mu}_{c,k}\|^2),$$

with

$$\mathrm{Tr}(\overline{\Sigma}_M) = C^{-1} \sum_{c,k} \pi_{c,k}\, \|\overline{\mu}_{c,k} - \overline{\mu}_c\|^2.$$

Inserting (12) gives

$$\mathrm{Tr}(\overline{\Sigma}_W) = \mathrm{Tr}(\overline{\Sigma}_M) + O(\sigma^{2-1/s} \log d). \tag{13}$$

By decomposing and inserting (11) we get

$$\begin{aligned}
\mathrm{Tr}(\overline{\Sigma}_M) &\leq C^{-1} \sum_{c,k} \pi_{c,k} \Big(\|\overline{\mu}_{c,k} - \mu_{c,k}\| + \|\mu_{c,k} - \mu_c\| + \|\mu_c - \overline{\mu}_c\|\Big)^2 \\
&= C^{-1} \sum_{c,k} \pi_{c,k} \Big(\|\mu_{c,k} - \mu_c\| + O(\sigma^{1-1/(2s)} \log^{1/2} d)\Big)^2 \\
&= C^{-1} \sum_{c,k} \pi_{c,k}\, 2\Big(\|\mu_{c,k} - \mu_c\|^2 + O(\sigma^{2-1/s} \log d)\Big) \\
&= 2\,\mathrm{Tr}(\Sigma_M) + O(\sigma^{2-1/s} \log d).
\end{aligned}$$

Inserting this inequality in (13) proves that

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

which proves (4).

We now prove Lemma B.1. Donoho & Johnstone (1994) proved that for all $d \geq 4$,

$$\mathbb{E}(\|\rho_t(x) - \mu\|^2) \leq (2 \log d + 1)\,(\sigma^2 + \sum_{m=1}^{d} \min(\mu[m]^2, \sigma^2)). \tag{14}$$

We are now going to prove that if $|\mu^{(r)}| \sim r^{-s}$ then

$$\sum_{m=1}^{d} \min(\mu[m]^2, \sigma^2) = O(\sigma^{2-1/s}).$$

Let us first observe that

$$\sum_{m=1}^{d} \min(\mu[m]^2, \sigma^2) = \sum_{r=M+1}^{d} |\mu^{(r)}|^2 + M\sigma^2 \tag{15}$$

with $|\mu^{(M)}| \geq \sigma > |\mu^{(M+1)}|$.

Since $|\mu^{(r)}| \sim r^{-s}$,

$$\sum_{m=1}^{d} \min(\mu[m]^2, \sigma^2) \sim \sum_{r=M+1}^{d} r^{-2s} + M\sigma^2 \sim M^{1-2s} + M\sigma^2.$$

Since $\sigma \sim |\mu^{(M)}| \sim M^{-s}$, we conclude

$$\sum_{m=1}^{d} \min(\mu[m]^2, \sigma^2) = O(\sigma^{2-1/s}).$$

Inserting this result in (14) finishes the proof of the lemma.

C Proof of Theorem 2.3

We choose $x = ru$ with $u \sim \mathcal{U}(\mathbb{S}^{d-1})$ and $r \in]0, 1]$ to be determined, with r and u independent. Let us fix $p \geq d$, $F \in \mathbb{R}^{p \times d}$, $W \in \mathbb{R}^{1 \times p}$ and $b \in \mathbb{R}$. With $g(x) = W\rho_{rt}Fx + b$, we have:

$$\begin{aligned} g(x) &= \sum_{m=1}^{p} w_m \rho_r(r\langle u, f_m\rangle - \lambda) + b \\ &= r \sum_{m=1}^{p} w_m \rho_r(\langle u, f_m\rangle - \lambda/r) + b\,. \end{aligned}$$

If $\lambda = 0$, this gives $g(x) = rW\rho_r(Fu) + b$ which is an affine function of r. Therefore, its sign can change at most once. We choose $h(x) = \cos(2\pi\|x\|)$ so that:

$$\operatorname{sgn}(h(x)) = \begin{cases} +1 & r < \frac{1}{4} \text{ or } \frac{3}{4} < r \\ -1 & \frac{1}{4} < r < \frac{3}{4} \end{cases}$$

Now $g(x)$ is an affine function of r, so at least one of the following must occur:

$$\begin{cases} \operatorname{sgn}(g(x)) = -1 & r < \frac{1}{4} \\ \operatorname{sgn}(g(x)) = +1 & \frac{1}{4} < r < \frac{3}{4} \\ \operatorname{sgn}(g(x)) = -1 & \frac{3}{4} < r \end{cases}$$

We finally choose $r \sim \mathcal{U}(0, 1)$ and so we conclude that:

$$\mathbb{P}[\operatorname{sgn}(g(x)) \neq \operatorname{sgn}(h(x))] \geq \frac{1}{4}\,.$$

If $\lambda > 0$, then when $r \leq \lambda$, we have $\langle u, f_m\rangle \leq \|u\|\|f_m\| \leq 1 \leq \lambda/r$, which means that $g(x) = b$ is constant. We thus choose $r \sim \mathcal{U}(0, \lambda)$, $h(x) = \cos(\pi/\lambda\|x\|)$ and so we conclude that:

$$\mathbb{P}[\operatorname{sgn}(g(x)) \neq \operatorname{sgn}(h(x))] = \frac{1}{2} \geq \frac{1}{4}\,.$$

D Implementation and Network Dimensions

All networks are trained with SGD with a momentum of 0.9 and a weight decay of 10^{-4} for the classifier weights, with no weight decay being applied to tight frames. The learning rate is set to 0.01 for all networks, with a Parseval regularization parameter $\alpha = 0.0005$. The batch size is 128 for all experiments. The scattering transform is based on the *Kymatio* package (Andreux et al., 2020).

Table 4: Number of parameters of scattering architectures on ImageNet. They are dominated by the size of the 1×1 orthogonal projectors P_j. Indeed, the wavelet tight frame F_w has a redundancy of $(L + 1/4)$, whereas in ResNet strided convolutions have a redundancy of $1/2$. This is due to the fact that F_w is not learned. However, F_w comes with a known structure across channels, which is beneficial for the analysis of the projectors P_j.

	Φ	S_T	S_P	S_C	ResNet-18
ImageNet	Parameters	25.9M	27.6M	31.2M	11.7M

Standard data augmentation was used on CIFAR and ImageNet: horizontal flips and random crops for CIFAR, and random resized crops of size 224 and horizontal flips for ImageNet. Classification error on ImageNet validation set is computed on a single center-crop of size 224.

Non-linearity thresholds are set to $\lambda = 1.5\sqrt{d/p}$ for the soft-thresholding ρ_t, and $\lambda = \sqrt{d/p}$ for the thresholded rectifier ρ_{rt}. Here d and p represent the dimension of the patches the convolutional operators F and F^T act on. To ensure that the fixed threshold is well adapted to the scale of the input x, we normalize all its patches so that they have a norm of $\sqrt{d}$. For 1×1 convolutional operators as in S_C, this amounts to normalizing the channel vectors at each spatial location in x.

Two-layer networks When learning a frame contraction directly on the input image, F is a convolutional operator over image patches of size $k \times k$ with a stride of $k/2$, where $k = 14$ for MNIST ($d = k^2 = 196$) and $k = 8$ for CIFAR ($d = 3k^2 = 192$). The frame F has p output channels, where $p = 2048$ for MNIST and $p = 8192$ for CIFAR. It thus maps each patch of dimension d to a channel vector of size $p \geq d$. Training lasts for 300 epochs, the learning rate being divided by 10 every 70 epochs.

Scattering tree We use $J = 3$ for MNIST and CIFAR and $J = 4$ for ImageNet. Each F_w uses $L = 8$ angles. It is followed by a standardization which sets the mean and variance of every channel to 0 and 1. We then learn a 1×1 convolutional orthogonal projector P_J to reduce the number of channels to $d = 512$. We finally apply a 1×1 spatial normalization, as before a tight frame thresholding. Training lasts for 300 epochs for MNIST and CIFAR (200 epochs for ImageNet), the learning rate being divided by 10 every 70 epochs (60 epochs for ImageNet).

Learned scattering We use $J = 4$ for CIFAR and $J = 6$ for ImageNet. Each F_w uses $L = 8$ angles. Each P_j is an orthogonal projector which is a 1×1 convolution. It reduces the number of channels to d_j with $d_1 = 64$, $d_2 = 128$, $d_3 = 256$ and $d_4 = 512$. For ImageNet, we also have $d_5 = d_6 = 512$. It is followed by a normalization which sets the norm across channels of each spatial position to $\sqrt{d_j}$. F_j is a 1×1 convolutional tight frame with p_j output channels, where $p_1 = 1024$, $p_2 = 2048$, $p_3 = 4096$ and $p_4 = 8192$ for CIFAR, $p_1 = 512$, $p_2 = p_3 = 1024$ and $p_4 = p_5 = p_6 = 2048$ for ImageNet. Training lasts for 300 epochs for CIFAR (200 epochs for ImageNet), the learning rate being divided by 10 every 70 epochs (60 epochs for ImageNet).

Fisher ratios Fisher ratios (eq. (1)) were computed using estimations of Σ_W and μ_c on the validation set. These estimations are unstable when the dimension d becomes large with respect to the number of data samples. To mitigate this, the Fisher ratios across layers from Table 3 were computed on the train set. Fisher ratios on ImageNet from Table 2 were computed only across channels, by considering each pixel as a distinct sample of the same class, in order to reduce dimensionality.

Figure 1: Examples of filters f_m from the convolutional tight frame F learned directly on the input x for CIFAR-10, using an absolute value non-linearity ρ_a. They resemble wavelet filters.