A LEARNING APPROACH FOR WAVELET DESIGN

Dhruv Jawali[†1]*, Abhishek Kumar*[†2] *and Chandra Sekhar Seelamantula*[3]

[1]National Mathematics Initiative, [2,3]Department of Electrical Engineering
Indian Institute of Science, Bangalore - 560012, India
Email: {dhruv13, kabhishek}@iisc.ac.in, chandra.sekhar@ieee.org

ABSTRACT

Wavelet analysis and perfect reconstruction filterbanks (PRFBs) are closely related. Desired properties on the wavelet could be translated to equivalent properties on a PRFB. We propose a new learning-based approach towards designing compactly supported orthonormal wavelets with a specified number of vanishing moments. We view PRFBs as a special class of *convolutional autoencoders*, which places the problem of wavelet/PRFB design within a learning framework. One could then deploy several state-of-the-art deep learning tools to solve the design problem. The PRFBs are learned by minimizing a squared-error loss function using gradient-descent optimization. The model is trained using a dataset containing random samples drawn from the standard normal distribution. We demonstrate that imposing orthonormality and vanishing moment constraints in the learning framework gives rise to filters that generate an orthonormal wavelet basis. We present results for learning PRFBs with filter lengths 2 and 8. As an illustration, we show that the proposed framework is able to learn the Daubechies wavelet with four vanishing moments, as well as wavelets with an arbitrary number of vanishing moments. For all our results, the signal-to-reconstruction error ratio is greater than 200 dB, implying that perfect reconstruction is indeed achieved accurately up to machine precision.

Index Terms— Wavelet design, autoencoders, filterbank learning, multiresolution analysis, vanishing moments

1. INTRODUCTION

Multiresolution analysis (MRA) has been extensively used for feature extraction from signals such as electrocardiograms (ECG) [1], electroencephalograms (EEG) [2] and images [3, 4]. The core idea is to capture signal structures across multiple scales by considering a dictionary that is composed of shifted and scaled versions of a single generator function $\psi(t)$:

$$\mathcal{D} = \left\{ \psi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j n}{2^j} \right) \right\}_{(n,j)\in\mathbb{Z}^2}.$$

If the function $\psi(t) \in \mathbf{L}^2(\mathbb{R})$ has a zero average and satisfies the admissibility criterion given by Calderón [5], Grossman and Morlet [6], it qualifies as a wavelet for performing the continuous-time wavelet transform. In a multiresolution analysis, wavelets enable a representation of the details at various scales.

A wavelet is said to have p vanishing moments if the following holds: $\int_{-\infty}^{\infty} t^k \psi(t) dt = 0$ for $k = 0, 1, 2, \ldots, p-1$. Wavelets having p vanishing moments annihilate polynomials up to order $p-1$ [7]. Thus, more the number of vanishing moments, sparser are the representations of regular signals. The sparsifying property makes wavelets useful in applications such as denoising and compression [8–12].

A real-world application of wavelets involves choosing one among many families of analytically derived wavelets based on properties such as regularity, number of vanishing moments, compact support, symmetry, and ease of implementation. Designing wavelets with compact support is typically achieved by designing a corresponding PRFB where the filters are constrained to satisfy the desired properties of the wavelet [13]. We transform the design problem to a learning problem so as to leverage state-of-the-art optimization tools and frameworks that have caused the *deep learning revolution* [14].

Learning based frameworks have been employed to address the problem of sparse representation of data. Pfister and Bresler relate learning sparsifying transforms to designing multidimensional, multirate filterbanks [15, 16]. They propose using gradient descent to learn the filters [16, 17]. The authors show applications of their approach on image denoising and accelerated magnetic resonance imaging. Using sparsity of the representation in wavelet bases and wavelet frames as a criterion to learn wavelets has been considered in [18] and [19], respectively. Recoskie and Mann represent the discrete wavelet transform as an autoencoder, and construct a loss function to impose sparsity in the learnt representation [18]. A drawback of their approach is that the learnt functions are not guaranteed to satisfy the properties of a wavelet. Tai and E search for a wavelet frame that perfectly reconstructs a given dataset while achieving a maximally sparse representation [19].

1.1. This Paper

In the present work, we view a two-channel PRFB as an autoencoder. The design problem is indirectly solved by training the autoencoder in a data-driven fashion subject to perfect reconstruction as the loss function. We also formulate the filterbank architecture to incorporate additional constraints based on the conjugate mirror filter property and vanishing moments property, which implicitly enforces sparsity in the associated representation. Since the perfect reconstruction property must be satisfied for any input sequence, the network must see a wide spectrum of inputs. For this purpose, we use random white Gaussian vectors as training data. The filters learnt are not specific to the data – the data essentially performs the job of a scaffolding to steer the optimization objective. Further, our experiments show that the proposed method indeed learns *bona fide* wavelets.

Getting to the specifics, we address the problem of learning orthonormal wavelet bases $\{\psi_{n,j}(t)\}_{(n,j)\in\mathbb{Z}^2}$, where $\psi(t)$ has a compact support $[0, L-1]$ and $p \leq \lfloor \frac{L}{2} \rfloor$ vanishing moments. We em-

†: Both authors have contributed equally.

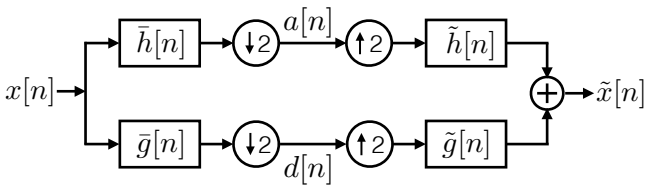

Fig. 1. A two-channel filterbank.

ploy a convolutional neural network (CNN) based autoencoder with the squared-error loss function. We propose a method to enforce p vanishing moments on the learnt wavelets. We show that the proposed learning framework is able to learn orthonormal wavelets such as the Haar wavelet and more generally the Daubechies family of wavelets accurately up to machine precision by imposing the constraint $L = 2p$. We also demonstrate that one could learn wavelets with an arbitrary number of vanishing moments by allowing $L \geq 2p$.

2. PERFECT RECONSTRUCTION FILTERBANKS

A two-channel multirate filterbank (Figure 1) splits an input signal $x[n]$ into two signals $a[n]$ and $d[n]$ having half the rate as $x[n]$. The analysis filters are denoted by $h[n] = \bar{h}[-n]$ and $g[n] = \bar{g}[-n]$, respectively, and $\tilde{h}[n]$ and $\tilde{g}[n]$ denote the synthesis filters. Vetterli [20] gave the following conditions on the four filters $h,\ \tilde{h},\ g$ and $\tilde{g}$ to reconstruct $x[n]$ exactly from $a[n]$ and $d[n]$:

$$\hat{h}^*(e^{\mathrm{j}(\omega+\pi)})\hat{\tilde{h}}(e^{\mathrm{j}\omega}) + \hat{g}^*(e^{\mathrm{j}(\omega+\pi)})\hat{\tilde{g}}(e^{\mathrm{j}\omega}) = 0, \tag{1}$$

$$\hat{h}^*(e^{\mathrm{j}\omega})\hat{\tilde{h}}(e^{\mathrm{j}\omega}) + \hat{g}^*(e^{\mathrm{j}\omega})\hat{\tilde{g}}(e^{\mathrm{j}\omega}) = 2, \tag{2}$$

where $\hat{h}(e^{\mathrm{j}\omega})$ is the discrete-time Fourier transform (DTFT) of $h[n]$. A collection of filters satisfying the above conditions is said to form a PRFB. For the case where the four filters have finite impulse response (FIR) with length L, the design problem reduces to solving an underdetermined system of linear equations. A common approach to deriving PRFBs analytically is to impose additional criteria to arrive at a unique solution. The following criteria are imposed on the filters to reduce the number of free variables to L:

$$\begin{aligned} &\tilde{h}[n] = h[n],\ \tilde{g}[n] = g[n] \text{ and} \\ &g[n] = (-1)^{1-n} h[(2l+1) - n],\ \text{where } l \in \mathbb{Z}. \end{aligned} \tag{3}$$

The filter h is then designed to satisfy (2), and is referred to as a conjugate mirror filter (CMF). Mallat [13] showed that, under certain conditions, such filterbanks are capable of performing MRA, the scaling (ϕ) and wavelet (ψ) functions of which are related to the filters h and g as

$$\hat{\phi}(\omega) = \prod_{p=1}^{\infty} \frac{\hat{h}\left(e^{\mathrm{j}2^{-p}\omega}\right)}{\sqrt{2}}; \quad \hat{\psi}(\omega) = \frac{1}{\sqrt{2}}\hat{g}\left(\frac{\omega}{2}\right)\hat{\phi}\left(\frac{\omega}{2}\right). \tag{4}$$

Lemarié showed that every compactly supported wavelet corresponds to an MRA [21]. Conversely, for a PRFB to generate a wavelet using (4), it can be verified that $\hat{g}(0) := \hat{g}(e^{\mathrm{j}\omega})|_{\omega=0} = 0$ must be satisfied, implying $\hat{h}(\pi) := \hat{h}(e^{\mathrm{j}\omega})|_{\omega=\pi} = 0$. Thus, the filter $h[n]$ must necessarily be lowpass.

If we further impose that $\psi(t)$ has p vanishing moments, $h[n]$ must have p roots at $\omega = \pi$ [22]. Daubechies [23] proved that a real filter $h[n]$ having p vanishing moments must have a minimum support of $2p$, giving rise to the so-called 'db$p$' family of wavelets having a support $2p$. Thus, it is possible to learn an orthonormal wavelet $\psi(t)$ having p vanishing moments by learning a filterbank with filter length $L \geq 2p$ by constraining the filter h to have p roots at $\omega = \pi$.

In the subsequent sections, we make use of the properties of the PRFB discussed so far in a learning framework.

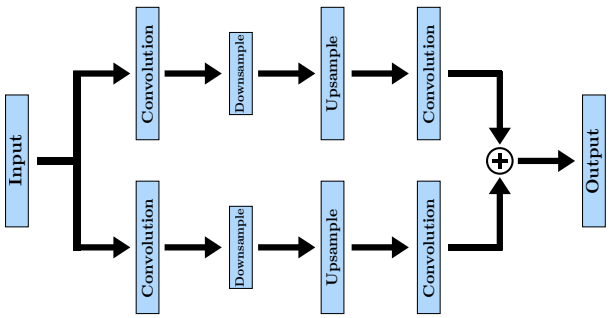

Fig. 2. The autoencoder architecture of a PRFB employed in this work.

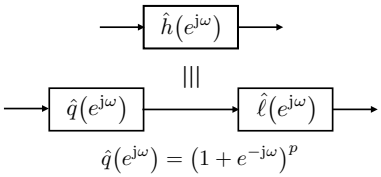

Fig. 3. Illustration of vanishing moments constraint on filter h. It is accomplished by imposing p roots at $\omega = \pi$.

3. VIEWING PRFBS AS CONVOLUTIONAL AUTOENCODERS

An autoencoder is a special type of a neural network that is trained to reproduce the given input at the output with a minimum distortion [24]. It is often used for learning lower-dimensional latent representations of the input data. In the case of a convolutional autoencoder [25], the convolutional layers are used for feature extraction followed by max-pool layers for subsampling.

A PRFB is essentially a convolutional autoencoder with the max-pool layers replaced by downsampling units. Max-pooling and downsampling are comparable in the objective they achieve – that of dimensionality reduction, but max-pooling is nonlinear whereas downsampling is a linear operation. Both are periodically shift-invariant. Owing to its linearity, the downsampling operation is easier to analyze than the max-pooling operation.

Our autoencoder shown in Fig. 2 has two parallel branches, each having separate layers performing convolution, downsampling, and upsampling operations. The same input is fed to both the branches, and the sum of their individual outputs is the final output.

The model is trained on a dataset $X = \{x_i \in \mathbb{R}^m\}_{1 \leq i \leq N}$, where x_is are independently drawn from the standard normal distribution. The data used for training needs to be generic enough to ensure perfect reconstruction for arbitrary signals from $\ell^2(\mathbb{Z})$. If the training data lies within a specific frequency band that is a subset of $[0, 2\pi]$, the learnt filters are likely to be frequency-selective and specific to that band, thus not generalizing well. Training is performed

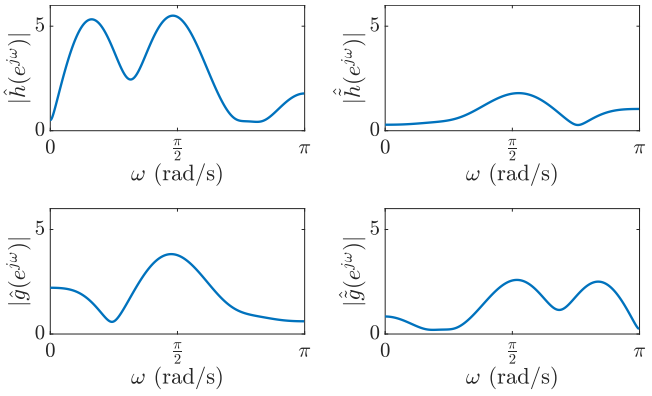

Fig. 4. Learning a PRFB with $L = 8$ and without imposing orthonormality or vanishing moments constraints. None of the filters learnt have frequency response that is zero at $\omega = 0$.

by minimizing the squared-error loss function:

$$\mathcal{L}(X; h, \tilde{h}, g, \tilde{g}) = \frac{1}{N} \sum_{i=1}^{N} \|x_i - \tilde{x}_i\|_2^2, \tag{5}$$

resulting in the optimization problem

$$h^*, \tilde{h}^*, g^*, \tilde{g}^* = \arg \min_{h,\tilde{h},g,\tilde{g}} \mathcal{L}(X; h, \tilde{h}, g, \tilde{g}). \tag{6}$$

This formulation is *unconstrained*, and requires learning a total of $4L$ parameters, corresponding to the four filters. However, one could also consider the constrained counterpart by imposing additional properties such as the CMF conditions (3), which requires L parameters to be learnt. We refer to imposing CMF conditions as the *orthonormality constrained* formulation.

Vanishing moments constraint: The vanishing moments of an orthonormal wavelet are tightly coupled to the zeros of the CMF h at $\omega = \pi$ (Theorem 7.4 of [22]) and the polynomials reproduced by the corresponding scaling function (*Strang-Fix* conditions, [7]). To enforce a desired number of vanishing moments, we introduce p roots at $\omega = \pi$ for h. This is accomplished by representing h as a cascade of two filters q and ℓ as shown in Figure 3. The first filter q has the frequency response $\hat{q}(e^{\mathrm{j}\omega}) = (1 + e^{-\mathrm{j}\omega})^p$, and a support of $p + 1$. The filter $\ell[n]$ is constrained to have a length of $L - p$, and is trained while $q[n]$ is kept fixed. With the constraint of p vanishing moments, the problem reduces to one of learning $L - p$ parameters instead of L.

In this paper, we consider the filters to be real-valued. However, one could also learn complex filters, but this would increase the search space dimension by a factor of 2.

4. IMPLEMENTATION

We implemented the PRFB learning using the TensorFlow Python library [26], which implements automatic differentiation to train the autoencoder, and is very convenient. We discuss the details regarding training, initialization, dataset, and performance metrics used in the proposed framework in this section.

4.1. Dataset and Initialization

We set the length m of each training signal x_i to 128 and use $N = 50$ signals for training. Our experiments showed that even a small dataset is adequate for learning a PRFB. The filter parameters were initialized as random samples drawn from a uniform distribution. It has been observed experimentally that initializing over a smaller support of the uniform distribution led to a faster convergence. We found that a range of $[-5, 5]$ was appropriate.

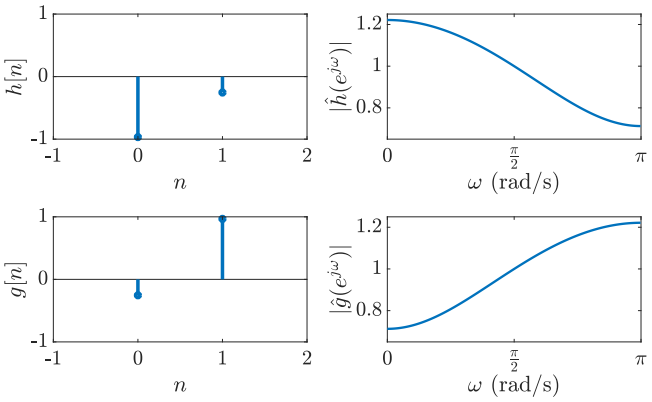

Fig. 5. Learning a PRFB with $L = 2$ and orthonormality constraints imposed.

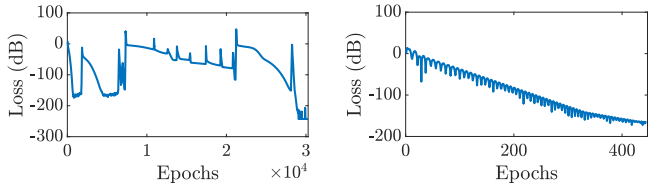

Fig. 6. Training loss plotted against epochs (a) corresponding to Figure 4; (b) corresponding to Figure 5.

4.2. Training

For training, we employ the gradient-descent-based Adam optimization algorithm [27]. The entire dataset X was used to compute the gradients at each iteration. The learning rate of the optimizer was tuned separately for each experiment to ensure convergence. We began with a high learning rate for each model, and reduced it whenever we observed large fluctuations in the training loss.

Training was continued until one of the following convergence criteria was satisfied: (i) the training loss going below a predefined threshold (10^{-28} in our case); or (ii) no change in the training loss for more than 100 iterations.

4.3. Performance Metric

We evaluate whether the filters learnt indeed form a PRFB in two ways. First, we compute the signal-to-reconstruction error ratio (SRER) defined as $\text{SRER} = 20 \log_{10} \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\|x_i\|_2}{\|x_i - \tilde{x}_i\|_2} \right)$ dB. A test set of white Gaussian noise signals $X_{\text{test}} = \left\{ x_i \in \mathbb{R}^{1000} \right\}_{1 \le i \le N}$ with $N = 100$ is used to compute the SRER. Observe that the dimensions of the training and test signals are different. This has been chosen so as to validate the perfect reconstruction property of the learnt filterbanks over the entire space of interest, $\ell^2(\mathbb{Z})$. Second, we numerically test whether the learnt filters satisfy the perfect

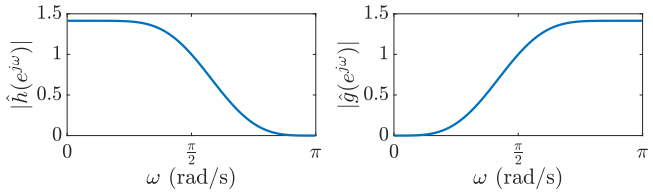

Fig. 7. Learning db4: Learnt PRFB filter responses with length $L = 8$, and $p = 4$ vanishing moments imposed.

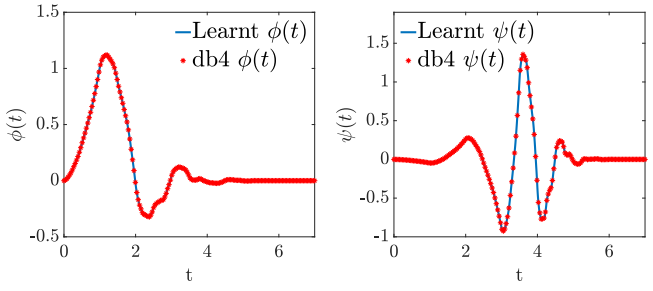

Fig. 8. [Colour online] The scaling and wavelet functions generated from the filters shown in Figure 7. The known db4 scaling and wavelet functions are superimposed.

reconstruction conditions (1) and (2) over a grid of frequencies in $[0, 2\pi]$.

For all the learnt filters reported in this paper, the SRER values turned out to be greater than 200 dB, and conditions (1) and (2) were satisfied with an accuracy up to machine precision (errors were of the order of 10^{-18}), implying that the proposed method is indeed able to learn PRFBs.

5. NUMERICAL RESULTS

5.1. Learning PRFBs

We first demonstrate learning PRFBs using the unconstrained as well as orthonormality constrained versions of our framework.

For the unconstrained case, we present the learnt filter responses for an $L = 8$ PRFB in Figure 4. We observe that the filters do not have the standard lowpass-highpass structure, which traditional filterbanks possess. None of the filters' frequency responses are zero at $\omega = 0$, implying that these do not generate true wavelet functions. This is an example of a PRFB that does not correspond to an MRA. Figure 6(a) shows the training loss as a function of epochs while learning the filterbank. We observe that though the training loss fluctuated several times, it did finally converge to a valid solution.

Next, we present results of a PRFB with the orthonormality constraint for filter length $L = 2$. The learnt filters and their frequency responses are presented in Figure 5. The training for this case completed within 500 epochs (cf. Figure 6(b)). It can be verified that for the length-two case, CMFs have an impulse response of the form $[\cos\theta, \sin\theta]$. The learnt filter has this form with $\theta = 255.26°$, thus agreeing well with the theory. In general, setting orthonormality constraints is not sufficient to guarantee that the learnt filters will generate a wavelet; we must additionally impose one root at $\omega = \pi$. For the $L = 2$ case, the Haar filterbank with impulse response $[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$ is the only possible CMF having a single root at $\omega = \pi$, corresponding to $\theta = 45°$.

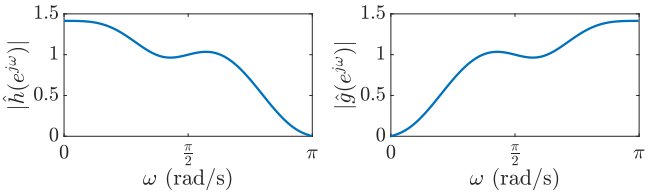

Fig. 9. Learning a wavelet with an arbitrary number of vanishing moments: Filters learnt with $L = 8$ and $p = 1$ vanishing moment.

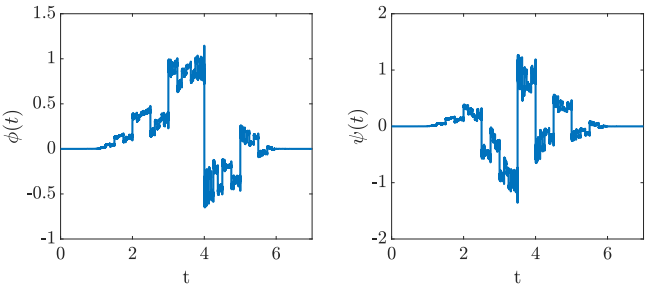

Fig. 10. Learning a wavelet with an arbitrary number of vanishing moments: scaling and wavelet functions obtained from learnt filters.

5.2. Learning wavelets with p vanishing moments

We next impose the vanishing moment constraint on the PRFB model. For filters of length $L = 8$, imposing $p = 4$ vanishing moments corresponds to the db4 wavelet. The frequency responses of the learnt filters are shown in Figure 7. The scaling and wavelet functions generated from the learnt filters are plotted in Figure 8. We also plot the db4 scaling and wavelet functions to facilitate comparison. The learnt scaling and wavelet functions match closely with the known db4 functions.

We conclude this section by showing that the proposed framework can learn wavelets with an arbitrary number of vanishing moments. We set $L = 8$ and $p = 1$. The learnt filters are presented in Figure 9. Imposing one vanishing moment ensures that an orthonormal wavelet basis is learnt, which can be verified by checking that $\hat{g}(0) = 0$. The learnt scaling and wavelet functions are shown in Figure 10. It is seen that the learnt wavelet is not smooth, which can be explained by Tchamitchian's theorem [28] (Chapter 7 in [22]). Tchamitchian established a relation between the Lipschitz regularity of the scaling and wavelet functions and the number of vanishing moments. To learn wavelets with a higher regularity, one can impose more vanishing moments and set the length accordingly.

6. CONCLUSIONS

We presented a learning based approach to arrive at wavelets with orthogonality and p vanishing moments criteria imposed on them. We showed that an autoencoder architecture having a structure similar to a two channel filterbank is able to learn filters satisfying the perfect reconstruction criteria. We described a method to impose p vanishing moments on wavelets by constraining the learnt filters to have p roots at $\omega = \pi$. We used this method to arrive at the well known db4 wavelet and showed that it is possible to learn an arbitrary wavelet having a single vanishing moment. Further, it is possible to impose additional criteria on the loss functions to learn wavelets adapted for special applications, a problem that requires more investigation.

7. REFERENCES

[1] C. Li, C. Zheng, and C. Tai, "Detection of ECG characteristic points using wavelet transforms," *IEEE Transactions on Biomedical Engineering*, vol. 42, no. 1, pp. 21–28, 1995.

[2] H. Adeli, Z. Zhou, and N. Dadmehr, "Analysis of EEG records in an epileptic patient using wavelet transform," *Journal of Neuroscience Methods*, vol. 123, no. 1, pp. 69–87, 2003.

[3] T. Chang and C.-C.J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Transactions on Image Processing*, vol. 2, no. 4, pp. 429–441, 1993.

[4] M. Unser, "Texture classification and segmentation using wavelet frames," *IEEE Transactions on Image Processing*, vol. 4, no. 11, pp. 1549–1560, 1995.

[5] A. Calderón, "Intermediate spaces and interpolation, the complex method," *Studia Mathematica*, vol. 24, no. 2, pp. 113–190, 1964.

[6] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM Journal on Mathematical Analysis*, vol. 15, no. 4, pp. 723–736, 1984.

[7] G. Strang and G. Fix, "A Fourier analysis of the finite element variational method," in *Constructive Aspects of Functional Analysis*, pp. 793–840. Springer, 2011.

[8] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[9] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE transactions on Image Processing*, vol. 9, no. 9, pp. 1532–1546, 2000.

[10] R. A. DeVore, B. Jawerth, and B. J. Lucier, "Image compression through wavelet transform coding," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 719–746, 1992.

[11] A. S. Lewis and G. Knowles, "Image compression using the 2-D wavelet transform," *IEEE Transactions on Image Processing*, vol. 1, no. 2, pp. 244–250, 1992.

[12] F. Luisier, T. Blu, and M. Unser, "A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding," *IEEE Transactions on Image Processing*, vol. 16, no. 3, pp. 593–606, 2007.

[13] S. Mallat, "Multiresolution approximations and wavelet orthonormal bases of $\mathbf{L}^2(\mathbb{R})$," *Transactions of the American Mathematical Society*, vol. 315, no. 1, pp. 69–87, 1989.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436, 2015.

[15] L. Pfister and Y. Bresler, "Learning sparsifying filter banks," in *Wavelets and Sparsity XVI*. International Society for Optics and Photonics, 2015, vol. 9597, p. 959703.

[16] L. Pfister and Y. Bresler, "Learning filter bank sparsifying transforms," *arXiv preprint arXiv:1803.01980*, 2018.

[17] L. Pfister and Y. Bresler, "Bounding multivariate trigonometric polynomials with applications to filter bank design," *arXiv preprint arXiv:1802.09588*, 2018.

[18] D. Recoskie and R. Mann, "Learning sparse wavelet representations," *arXiv preprint arXiv:1802.02961*, 2018.

[19] C. Tai and W. E, "Multiscale adaptive representation of signals: I. the basic framework," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 4875–4912, 2016.

[20] M. Vetterli, "Filter banks allowing perfect reconstruction," *Signal Process.*, vol. 10, no. 3, pp. 219–244, Apr. 1986.

[21] P. G. Lemarié, *Les ondelettes en 1989: séminaire d'analyse harmonique, Université de Paris-Sud, Orsay*, vol. 1438, Springer, 2006.

[22] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Academic press, 2008.

[23] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, pp. 909–996, 1988.

[24] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 37–49.

[25] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International Conference on Artificial Neural Networks*. Springer, 2011, pp. 52–59.

[26] M. Abadi et al., "Tensorflow: a system for large-scale machine learning.," in *OSDI*, 2016, vol. 16, pp. 265–283.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Preprint arXiv:1412.6980*, 2014.

[28] P. Tchamitchian, "Biorthogonalité et théorie des opérateurs," *Revista Matemática Iberoamericana*, vol. 3, no. 2, pp. 163–189, 1987.