SPARSE DEEP SCATTERING CROISÉ NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we propose the Sparse Deep Scattering Croisé Network (SDCSN) a novel architecture based on the Deep Scattering Network (DSN). The DSN is achieved by cascading wavelet transform convolutions with a complex modulus and a time-invariant operator. We extend this work by first, crossing multiple wavelet family transforms to increase the feature diversity while avoiding any learning. Thus providing a more informative latent representation and benefit from the development of highly specialized wavelet filters over the last decades. Beside, by combining all the different wavelet representations, we reduce the amount of prior information needed regarding the signals at hand. Secondly, we develop an optimal thresholding strategy for over-complete filter banks that regularizes the network and controls instabilities such as inherent non-stationary noise in the signal. Our systematic and principled solution sparsifies the latent representation of the network by acting as a local mask distinguishing between activity and noise. Thus, we propose to enhance the DSN by increasing the variance of the scattering coefficients representation as well as improve its robustness with respect to non-stationary noise. We show that our new approach is more robust and outperforms the DSN on a bird detection task.

1 INTRODUCTION

Modern Machine Learning focuses on developing algorithms to tackle natural machine perception tasks such as speech recognition, computer vision, recommendation among others. Historically, some of the proposed models were based on well-justified mathematical tools from signal processing such as Fourier analysis. Hand-crafted features were then computed based on those tools and a classifier was trained supervised for the task of interest. However, such theory-guided approaches have become almost obsolete with the growth of computational power and the advent of high-capacity models. As such, over the past decade the standard solution evolved around deep neural networks (DNNs). While providing state-of-the-art performance on many benchmarks, at least two pernicious problems still plague DNNs: First, the absence of stability in the DNN's input-output mapping. This has famously led to adversarial attacks where small perturbations of the input lead to dramatically different outputs. In addition, this lack of control manifests in the detection thresholds (i.e: ReLU bias) of DNNs, rendering them prone to instabilities when their inputs exhibit non-stationary noise and discontinuities. Second, when inputs have low SNR, or classes are unbalanced, the stability of DNNs is cantilevered. A common approach to tackle this difficulty is to increase both the size of the training set and the number of parameters of the network resulting in a longer training time and a costly labeling process. In order to alleviate these issues we propose the use of the DSN by creating a new non-linearity based on continuous wavelet thresholding. Thus our model, inherits the mathematical guarantees intrinsic to the DSN regarding the stability, and improves the control via wavelet thresholding method. Then, in order to produce time-frequency representation that are not biased toward a single wavelet family, we propose to combine diverse wavelet families throughout the network. Increasing the variability of the scattering coefficient, we improve the linearization capability of the DSN and reduce the need of an expert knowledge regarding the choice of specific filter bank with respect to each input signal.

The paper is organized as follows: 1.1 and 1.2 are devoted to the related work and contribution of the paper, the section 2 shows the theoretical results, where 2.1 is dedicated to the network architecture and its properties, and 2.2 provides the milestone of our thresholding method, then section 2.3 shows the characterization, via latent representation, of our network on different events by on

the Freefield1010[1] audio scenes dataset. Finally, we evaluate our architecture and compare it to the DSN on a bird detection task are shown in 2.4. The appendix in divided into three parts, Appendix A provides both, the pre-requisite and details about building the wavelets dictionary to create our architecture; Appendix B shows additional results on the sparsity of the SDCSN latent representations; Appendix C shows mathematical details and proofs for the over-complete thresholding non-linearity.

1.1 Related Work

We extend the Deep Scattering Network, first developed in Mallat (2012) and first successfully applied in Bruna & Mallat (2011); Andén & Mallat. The Scattering Network (SN) is a cascade of linear and non-linear operators on the input signal. The linear transformation is a wavelet transform, and the nonlinear transformation is a complex modulus. For each layer, the scattering coefficients are computed according to the application of the scaling function on the representation. This network is stable (Lipschitz-continuous) and suitable for machine learning tasks as it removes spatiotemporal nuisances by building space/time-invariant features. The translation invariant property is provided by the scaling function that acts as an averaging operator on each layer of the transform leading to an exponential decay of the scattering coefficients Waldspurger (2017). Since the continuous wavelet transform increases the number of features, the complex modulus is used as its contractive property reduces the variance of the projected space Mallat (2016). Two extensions of this architecture have been already developed: the Joint Scattering Network Andén et al. (2015) and the time-chroma-frequency scattering Lostanlen & Mallat (2016). They introduced extra parameterization of the wavelets coefficient in the second layer of the network to capture frequency correlations allowing the scattering coefficient to represent the transient structure of harmonic sounds.

Thresholding in the wavelet domain remains a powerful approach for signal denoising as it exploits the edge-detector property of wavelets, providing a sparse representation of the input signal in the time-frequency plane. This property is characterized for each wavelet by its vanishing moments expressing the orthogonality of the wavelet with respect to a given order of smoothness in the input signal. We base our approach on the theories relating the thresholding of signal in the wavelet basis and the evaluation of the best basis. Both are realized via a risk evaluation that arose from different perspectives: statistical signal processing Donoho et al.; 1995); Krim et al. (1999), information theory Coifman & Wickerhauser (1992); Wijaya et al. (2017); Cosentino et al. (2016), and signal processing Mallat & Zhang (1993); Mallat (1999). However, to the best of our knowledge, there is no thresholding method developed for continuous wavelet transform. We will thus extend the work of Berkner (1998) on thresholding over-complete dictionnary in the case of TIDWT and Biorthogonal-DWT to build a risk evaluation in the case of over-complete continuous wavelet transform.

1.2 Contributions

As opposed to the chroma-time-frequency scattering, using one wavelet family filter bank but deriving many symmetries of the latter, we propose to use multiple wavelet families having complementary properties (described in A.2) within a unified network yielding cross connections. It helps the architecture to provide higher dimensional and uncorrelated features, reducing the need of an expert to hand-choose the DSN wavelet filters, and also enables any downstream classifier to have greater orbit learning capacity. Therefore our architecture, the Deep Croisé Scattering Network (DCSN), leverages the simultaneous decomposition of complementary filter-banks as well as their crossed decomposition, hence the term "croisé." Then, endowing this architecture with our novel thresholding operator, we build the SDSCN providing new features based on the reconstruction risk of each wavelet dictionary. This method based on empirical risk minimization will bring several advantages. First, it enables us to insure and control the stability of the input-output mapping via thresholding the coefficients. Second, the model has sparse latent representations that ease the learning of decision boundaries as well as increases generalization performance. Finally, the risk associated with each wavelet family provides a characterization of the time-frequency components of the analyzed signal, that, when combines with scattering features enhances the high linearization capacity of DSN. As opposed to ReLU-based nonlinearities that impose sparsity by thresholding coefficients based on a

[1] http://machine-listening.eecs.qmul.ac.uk/bird-audio-detection-challenge/

fixed learned scalar threshold, we propose an *input-dependant locally adaptive* thresholding method. Therefore, our contribution leading to the Sparse Deep Croisé Network is twofold:

- Deep Croisé Scattering Network: a natural extension of the DSN allowing the use of multiple wavelet families and their crossed representations.
- Derivation of optimal non-orthogonal thresholding for overcomplete dictionaries: empirical risk minimization leads to an analytical solution for the denoising mask, allowing deterministic per-input solutions, and endowing the DSN with sparse latent representations.

2 SPARSE DEEP CROISÉ SCATTERING NETWORK

The Deep Croisé Scattering Network (DCSN) is a tree architecture (2 layers of such model are shown in Fig. 1) based on the Deep Scattering Network (DSN). The first layer of a scattering transform corresponding to standard scalogram is now replaced with a $3D$ tensor by adding the wavelet family dimension. Hence, it can be seen as a stacked version of multiple scalograms, one per wavelet family. The second layer of the DCSN brings inter and intra wavelet family decompositions. In fact, each wavelet family of the second layer will be applied on all the first layer scalograms, the same process is successively applied for building deeper model.

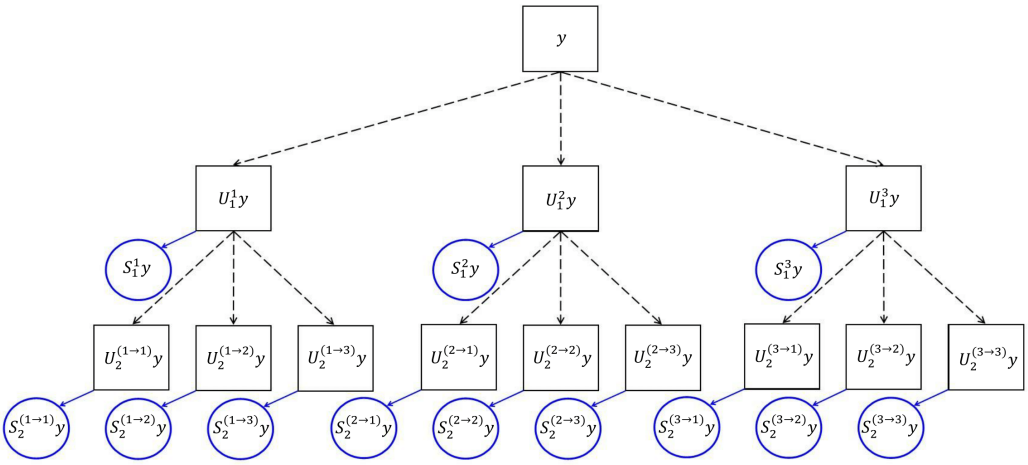


Figure 1: Deep Croisé Scattering Network Architecture for an input signal y - Special case of 3 wavelet families.

2.1 CROISE SCATTERING NETWORK: CROSS FAMILY REPRESENTATION FOR SIGNAL CHARACTERIZATION

We first proceed by describing the formalism of the DCSN $\forall x \in L^2$, details on wavelets and filter bank design are provided in Appendix A. We denote by

$$\Psi^{(1)} = \{\psi^{(1,b)}, b = 1, \dots, B^{(1)}\}, \tag{1}$$

the collection of $B^{(1)}$ mother wavelets for the first layer. We also denote by,

$$\lambda^{(1)} = \left\{\lambda_j^{(1)}, j = 1, ..., J^{(1)} \times Q^{(1)}\right\}, \quad \lambda_j^{(1)} = 2^{(j-1)/Q^{(1)}}, \tag{2}$$

the resolution coefficients for this first layer with $J^{(1)}$ representing the number of octave to decompose and $Q^{(1)}$ the quality coefficients a.k.a the number of wavelets per octave. Based on those configuration coefficients, the filter-banks can be derived by scaling of the mother wavelets through

the resolution coefficients. We thus denote the filter-bank creation operator $\mathcal{W}$ by

$$\mathcal{W}[\psi^{(1,b)}, \lambda^{(1)}] = \begin{pmatrix} \psi^{(1,b)}_{\lambda^{(1)}_1} \\ \cdots \\ \psi^{(1,b)}_{\lambda^{(1)}_{J^{(1)}\times Q^{(1)}}} \end{pmatrix}, \quad \psi^{(1,b)}_{\lambda^{(1)}_j}(t) = \frac{1}{\sqrt{\lambda^{(1)}_j}} \psi^{(1,b)} \left(\frac{t}{\lambda^{(1)}_j} \right). \tag{3}$$

To avoid redundant notation, we thus denote this filter-bank as $\mathcal{W}^{(1,b)}$ with implicit parameters $\Psi^{(1)}$ and $\Lambda^{(1)}$. We now developed of the needed tools to explicit define the filter layer of the DCSN. We denote by $U^{(1)}$ the output of this first layer and as previously mentioned it consist of a $3D$ tensor of shape $(B^{(1)}, J^{(1)}Q^{(1)}, N)$ with N the length of the input signal denoted as x. We omit here boundary conditions, sub-sampling, and consider a constant shape of N throughout the representations. We thus obtain

$$U^{(1)}[x](b, ., .) = |x \star \mathcal{W}^{(1,b)}|, b = 1, \ldots, B^{(1)}, \tag{4}$$

where $|.|$ operator corresponds to an element-wise complex modulus application. We define the convolution operation between those two objects as

$$x \star \mathcal{W}^{(1,b)} = \begin{pmatrix} x \star \mathcal{W}^{(1,b)}(1, .) \\ \cdots \\ x \star \mathcal{W}^{(1,b)}(J^{(1)}Q^{(1)}, .) \end{pmatrix}. \tag{5}$$

From this, the second layer we present below will introduce the cross family representations. First, we denote by $\lambda^{(2)}$ and $\Psi^{(2)}$ the internal parameters of layer 2 analogous to the first layer definition. We now denote the second layer representation as $U^{(2)}[U^{(1)}[x]]$. This object is a $5D$ tensor introduced 2 extra dimension on the previous tensor shape. In fact, is it defined as

$$U^{(2)}[U^{(1)}[x]](b_2, j_2, b_1, ., .) = U^{(1)}(b_1, ., .) \star \mathcal{W}^{(\ell, b_2)}(j_2)|, b_2 = 1, \ldots, B^{(2)}. \tag{6}$$

from this, we denote by croisé representation all the terms in $U^{(2)}[U^{(1)}[x]](b_2, j_2, b_1, ., .)$ with $b_2 \neq b_1$. For notations clarity we denote this representation as $U^{(2)}[U^{(1)}[x]](b_2, j_2, b_1, ., .) := U^{b_1 \to b_2}_{j_2}[x]$. Based on those notations it is straightforward to extend those representation to layer ℓ as $U^{b_1 \to \cdots \to b_\ell}_{j_2, \ldots, j_\ell}[x]$. We however limit ourselves in practice to 2 layers as usually done with the standard scattering networks. Given those representations, the scattering coefficients, the features per say, are defined as follows:

$$S^{b_1 \to \cdots \to b_\ell}_{j_2, \ldots, j_\ell}[x] = U^{b_1 \to \cdots \to b_\ell}_{j_2, \ldots, j_\ell}[x] \star \phi, \tag{7}$$

with ϕ is a scaling function. This application of a low frequency band-pass filter allows for symmetries invariances, inversely proportional to the cut-off frequency of ψ. We present an illustration of the network computation in Fig. 2.

As can be seen in the proposed example, while the first layer provides time-frequency information, the second layer characterizes transients as demonstrated in Balestriero & Aazhang (2016). With this extended framework, we now dive into the problem of thresholding over complete basis, cases where the quality factor, Q, is greater than 1 which are in practice needed to bring enough frequency precision.

2.2 Sparsity and Winner-take-all via Risk Minimization: Optimal Overcomplete Basis Thresholding

Sparsity in the latent representation of connectivists models have been praised many times Narang et al. (2017); Liu et al. (2015); Thom & Palm (2013). It represents the fitness of the internal parameters of a model needed with only few nonzeros coefficients to perform the task at hand. Furthermore, sparsity is synonym of simpler models as directly related with the Minimum Description Length Dhillon et al. (2011) guaranteeing increased generalization performances. In addition of those concepts, thresholding brings in practice robustness to noise. In particular, as we will demonstrate, even in large scale configuration, non-stationnary noise can not be completely handled by common ML approaches on their own. To do so, we propose to extend standard wavelet thresholding techniques for non-orthogonal filter-banks. Our approach aims at minimizing the reconstruction error of the thresholded signal in the wavelet domain via an oracle decision. Through this formulation, we are able to derive an analytical thresholding based on the input representation and the filter-bank redundancy. We now propose to derive this scheme and then provide interpretation on its underlying tests and computations.

2.2.1 IDEAL RISK AND EMPIRICAL RISK BOUND

As the decomposition is not orthogonal, the first point to be tackle is the difference of the L^2 approximation errors in between the original basis and the over-complete wavelet basis as Parseval equality does not hold. Beside, the transpose of the change of basis matrix is not anymore the inverse transform. Berknet et. al. in Berkner (1998) proposed the use of the Moore pseudo inverse to build the reconstruction dictionary. In the following we develop an upper bound to the ideal risk such that we benefit an explicit equation for the thresholding operator that is adapted to any over-complete transformation. Let's assume the observed signal, denoted by y, is corrupted with white noise such that $y = x+\epsilon$ where x is the signal of interest and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We now denote by $W \in \mathbb{C}^{N(J*Q+1)\times n}$ the matrix composed by the the wavelets at each time and scale (i.e: explicit convolution version of $\mathcal{W}$) such that $\forall x \in \mathbb{R}^N$, Wx is the wavelet transform. We denote by $W^\dagger \in \mathbb{C}^{n\times n(J*Q+1)}$ the generalized inverse such that $W^\dagger W = I$. The estimate of x is given by $\hat{x}_{W,D}(y) = W^\dagger D^S W y$.

$$\mathcal{R}^\star(x, W) = \min_\delta \mathbb{E}\left\|x - \hat{x}_{W,D}(y)\right\|^2 = \min_\delta \mathbb{E}\left\|W^\dagger(Wx - D^S W y\right\|^2 \tag{8}$$

Because of the correlation implied by the redundant information contained in the filter banks, the ideal risk is now dependent on all the possible pairs in the frequency axis. However,the independence in time remains. Since this optimization problem does not have an analytic expression, we propose the following upper bound explicitly derived in Appendix C.1. The upper-bound on the optimal risk is denoted by $\mathcal{R}_{up}$ and defined as,

$$\mathcal{R}_{up}(x, W) = \sum_{k=1}^{n(J*Q+1)} \min(\mathcal{R}^U_{up}(x), \mathcal{R}^S_{up}). \tag{9}$$

where we denote by $\mathcal{R}^U_{up}$ the upper bound error term corresponding to unselected coefficients:

$$\mathcal{R}^U_{up}(x) = \sum_{j=1}^{n*(J*Q+1)} \left|\mu_k(x)\mu j(x) \sum_{t=1}^{n} \psi_t^\dagger[k]\psi_t^\dagger[j]\right|, \tag{10}$$

and by $\mathcal{R}^S_{up}$ the upper bound error term corresponding to the selected coefficients:

$$\mathcal{R}^S_{up} = \sigma^2 \sum_{j=1}^{n(J*Q+1)} \left|\sum_{t=1}^{n} (\psi_t^\dagger[k]\psi_t^\dagger[j])\psi_k^T\psi_j\right|. \tag{11}$$

Now, one way to evaluate this upper-bound is to assume an orthogonal basis, and to compare it with the optimal risk in the orthogonal case which leads to the following proposition.

Proposition 1. *Assuming orthogonal filter matrix W_O, the upper bound ideal risk coincides with the orthogonal ideal risk:*

$$\mathcal{R}_{up}(x, W_O) = \mathcal{R}_O(x, W_O)$$

the proof is derived in C.2 In order to apply the ideal risk derive, ones needs an oracle decision regarding the signal of interest. In real application, the signal of interest x is unknown. We thus propose the following empirical risk:

$$\tilde{\mathcal{R}}_(y, W) = \sum_{k=1}^{n(J*Q+1)} \min(\mathcal{R}^U_{up}(y), \mathcal{R}^S_{up}). \tag{12}$$

This risk corresponds to the empirical version of the ideal risk where the observed signal y is evaluate in the left part of the minimization function. In order to compare this empirical risk with the ideal version, we propose their comparison the following extreme cases:

Proposition 2. *In the case where $D^S = I$, the empirical risk coincides with the upper bound ideal risk:*

$$\tilde{\mathcal{R}}(y, W) = \mathcal{R}_{up}(x, W).$$

Proposition 3. *In the case where $D^U = I$, the following bound shows the distance between the empirical risk and the upper bound ideal risk:*

$$\tilde{\mathcal{R}}(y, W) \leq \mathcal{R}_{up}(x, W) + C \times \left|\sum_{t=1}^{n} \psi_t^\dagger[k]\psi_t^\dagger[j]\right|, a.s. \tag{13}$$

where,

$$C = \sum_{k=1}^{n(J*Q+1)} \sum_{j=1}^{n(J*Q+1)} |\mu_k(x)| \, \|\psi_j\|_1 \, \sigma\sqrt{\frac{2}{\pi}} + |\mu_j(x)| \, \|\psi_k\|_1 \, \sigma\sqrt{\frac{2}{\pi}} + \sigma^2(1 - \frac{2}{\pi}).$$

Refer to C.3 for proofs. As the empirical risk introduces the noise in the left part of the risk expression, this term represents the propagation of this noise throughout the decomposition. We provided a generic development of the risk minimization process. When applied to a particular path of the scattering network, it is denoted as,

$$R^{b_1 \to \cdots \to b_\ell}_{j_1,\ldots,j_{\ell-1}}[y] = \mathcal{R}[U^{(\ell)}[y](b_\ell, ., b_{\ell-1}, j_{\ell-1}, \ldots, b_1, j_1, .), \mathcal{W}^{(\ell,b_\ell)}], b_\ell = 1, \ldots, B^{(\ell)}, \tag{14}$$

with $U^{(\ell)}[y](b_\ell, ., b_{\ell-1}, j_{\ell-1}, \ldots, b_1, j_1, .) \in \mathbb{R}^{J^{(\ell)}Q^{(\ell)} \times N}$ and $\mathcal{R}$ representing the risk minimization operator based on a given representation and the associated filter-bank.

2.2.2 Deep Croisé Scattering Network Thresholding Operator

We define by $\mathcal{T}$ the tresholding operator minimizing the the empirical risk,

$$\mathcal{T} = \arg\min_{\delta} \tilde{\mathcal{R}}(y, W). \tag{15}$$

In particular when applied to a specific path of the tree, this thresholding operator is denoted as $\mathcal{T}[U^{(\ell)}[y](b_\ell, ., b_{\ell-1}, j_{\ell-1}, \ldots, b_1, j_1, .), \mathcal{W}^{(\ell,b_\ell)}], b_\ell = 1, \ldots, B^{(\ell)}$. We provide in Fig. 2 illustration showing the effect of this thresholding operator at each layer of the network.

2.3 Risk Based Filter-banks Fitness Evaluation for Structural Event Characterization

We demonstrated in the last section the important of the risk in the optimal thresholding optimization. This empirical version of this risk represents the ability of the induced representation to perform efficient denoising and signal reconstruction. This concept is identical to the one of function fitness when considering the denoised ideal signal x and the thresholded reconstruction. As a result, it is clear that the optimal basis given a signal is the one with minimal empirical risk. We thus propose here simple visualization and discussion on this concept and motivate the need to use the optimal empirical risk as part of the features characterizing the input signal y along all the representations.

In Fig. 3, we provide two samples from the dataset corresponding to very different acoustic scene. One represents transients on the right while the left one provides mixture of natural sounds. Risk based analysis of the filter-banks fitness provide consistent information with the specificities of the selected wavelets. In fact, Paul family is known to be very adapted for transient characterization via its high time localization. On the other hand, the Morlet wavelet is optimal in term of Heisenberg principle and thus suitable for natural sounds such as bird songs, speech, music.

2.4 Validation: Bird Activity Detection benchmark

We propose to validate the two contributions over a large-scale audio dataset. As we will demonstrate below, our method as well as each contribution taken independently and jointly lead to significant increase in the final performance. We compare our results against the standard SN. In all cases, the scattering coefficients are then fed into a random forest Breiman (2001) with parameters[2] based on the sklearn library Pedregosa et al. (2011)

2.4.1 Challenge Overview and Standard Solutions

The data set consists of 4000 field recording signals from freefield1010[3] collected via the Freesound[4] project. This collection represent a wide range of audio scenes such as birdsong,city,nature,people,train,voice,water... The focus in this paper is the bird audio detection task

[2] n_estimator: 100, criteria: 'gini', min_samples_split: from 50 (small data size) to 150 (all data),class_weights:'balanced_subsample'

[3] http://machine-listening.eecs.qmul.ac.uk/bird-audio-detection-challenge/

[4] https://arxiv.org/abs/1309.5275

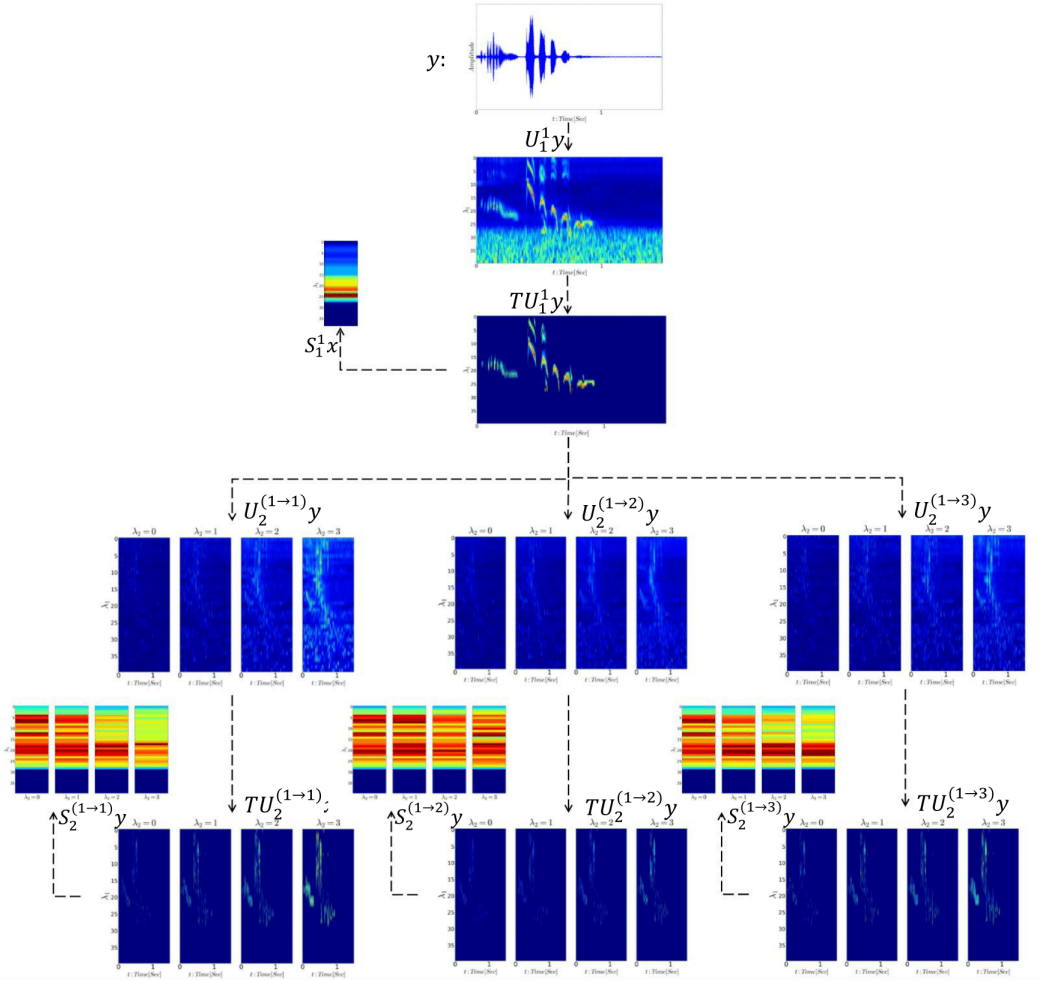
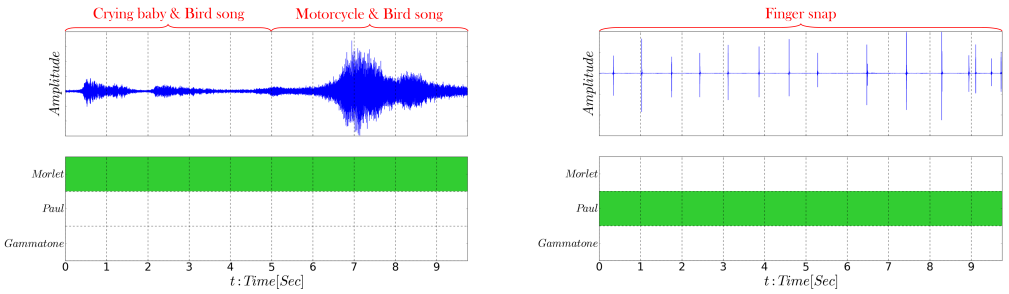


Figure 2: **Sparse Deep Croisé Scattering Network** representation - Paths: $1 \to 1 :, 1 \to 2, 1 \to 3$ with 1 : Morlet, 2 : Paul, 3: Gammatone - J1= 5, Q1=8, J2=4, Q2=1



(a) (b)

Figure 3: **Sparse Croisé Latent Wavelet Family Representation**: (a) Cocktail: Crying Baby & Bird song & Motorcycle - 15657.wav (c) Finger Snap Latent Representation - 349.wav

that can be formally defined as a binary classification task, where each label corresponds to the presence or absence of birds. Each signal is 10sec. long, and has been sampled at 44.1Khz. The evaluation of the results is performed via the Area Under Curve metric on 33% of the data. The experiments are repeated 50 times. The total audio length of this dataset is thus of slightly more than 11 hours of audio recordings. To put in comparison, it is about $10\times$ larger than CIFAR10 in term of

numbers of scalar values in the dataset. The results comparing our algorithm to the DSN with each of the wavelet family used in both SDCSN and DCSN are in Table 1. Both the SDCSN and DCSN outperform from at least 20% accuracy of any DSN proving the enhancement of the scattering feature by including the crossed latent representations. For all the architectures, the octave and quality parameters of the layers are $J1 = 5, Q1 = 8, J2 = 4, Q2 = 1$. As the feature of interests are birds songs, only high frequency content requires high resolution, the thresholding is applied per window of 2^{16} representing $\approx 1.5sec$.

Table 1: Classification Results - Bird Detection - Area Under Curve metric

	min	***mean***	***max***
Sparse Deep Croisé Scattering Network			
$S_1\ S_2\ R_1\ R_2$	**70.08**	**73.33**	**78.51**
$S_1\ R_1$	69.52	72.52	74.98
Deep Croisé Scattering Network			
$S_1\ S_2$	68.77	71.66	74.17
S_1	68.98	71.52	74.03
Deep Scattering Network			
Morlet - $S_1\ S_2$	51.88	53.57	55.62
Paul - $S_1\ S_2$	52.11	54.58	56.88
Gammatone - $S_1\ S_2$	51.34	54.11	56.09
Morlet - S_1	48.11	50.66	53.60
Paul - S_1	48.63	52.45	54.84
Gammatone - S_1	50.78	53.01	55.10

2.4.2 Proposed Solution and Controlled Denoising Experiment

We based our implementation on Balestriero & Glotin (2017) leveraging the Fourier based computations and localized filter in the frequency domain. For all input signals we perform a renormalization such that all inputs have unitary energy. We provide series of experiments, each demonstrating the benefits of our networks.

When considering different dataset sizes, the impact of denoising can be analyzed in details in Fig. 4a. As the dataset becomes smaller, the thresholding operator removing the noise perturbations becomes mandatory.. With infinite data and very high capacity classifier, a priori denoising becomes redundant as it is possible for the classifier to leverage the variance of the data to adjust correctly the hyperplanes delimiting the class regions. However, doing such learning is not possible with small scale dataset hence requiring a priori and deterministic denoising.

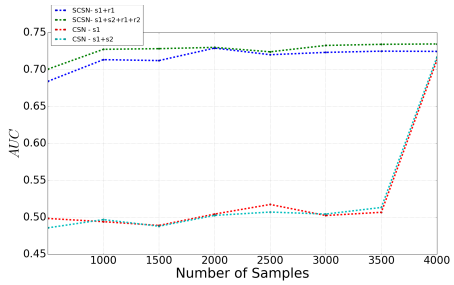
Another experiment highlighting the need for denoising in practical application comes from the possibility to have different noise levels from the training set to the test set. Thus we propose in Fig. 4b the following experiment. For both the SDCSN and DCSN models, training is achieved done on the denoised dataset. Then, the testing phase is performed on the raw dataset. Clearly, performances degrade strongly for the DCSN showing the inhability of the classifier, even though after standard scattering network transform, to be robust to noise level changes during and after training. This shows empirically the need of a thresholding non-linearity to provide more robustness to Scattering networks.

2.4.3 Sparsity Induced via Thresholding

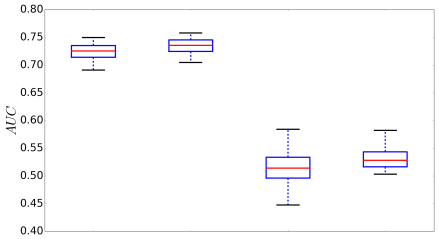
We now propose to visualize the sparsity induced via our thresholding technique (Fig. 5,10). To do so we present histograms of the representation with and without thresholding. Greater sparsity coupled with better performances and closely related to better linearization capacities, which benefits greatly the classifier as the size of the data is small 4a.

3 Conclusion and Future Work

We presented an extension of the scattering network so that one can leverage multiple wavelet families simultaneously. Via a specific topology, cross family representations are performed carrying

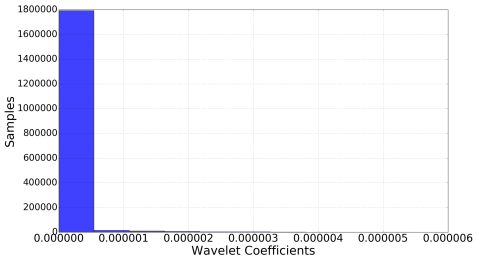


(a) **SDCSN and DCSN** - AUC Evaluation with respect to data-set size - Average over 50 runs - 33 % test set

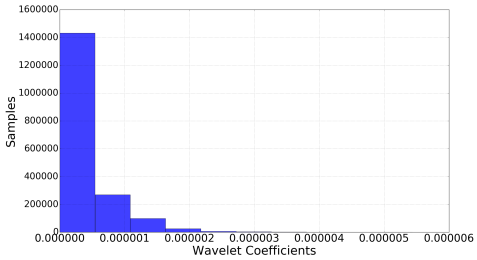


(b) **SDCSN and DCSN:** Training on denoised and Testing on noisy data vs denoised data: *From left to right*: - *SDCSN* S1+R1 & *SDCSN* S1+S2+R1+R2 train/test on denoised - *DCSN* S1 & *CSN* S1+S2 train on denoised/test on noisy

Figure 4: **Robustness Evaluation**



(a) **SDCSN** - Path: $1 \rightarrow 1$, 1:Gammatone



(b) **DCSN** - Path: $1 \rightarrow 1$, 1:Gammatone

Figure 5: **Histogram of activation of the first layer** - signal and decomposition are provided in 2

crucial information, as we demonstrated experimentally, allowing to significantly outperform standard scattering networks. We then motivated and proposed analytical derivation of an optimal overcomplete basis threhsolding being input adaptive. By providing greater sparsity in the representation but also a measure of filter-bank fitness. Again, we provided experimental validation of the use of our thresholding technique proving the robustness implied by such non-linearity. Finally, the ability to perform active denoising has been demonstrated crucial as we demonstrated that even in large scale setting, standard machine learning approach coupled with the SN fail to discard non-stationary noise. This coupled with the denoising ability of our approach should provide real world application the stability needed for consistent results and prediction control.

Among the possible extensions is the one adapting the technique to convolutional neural networks such that it provides robustness with respect to adversarial attacks. Furthermore, using a joint scattering and DNN will inherit the benefits presented with our technique as our layers are the ones closer to the input. Hence, denoising will benefit the inner layers, the unconstrained standard DNN layers. Finally, it is possible to perform more consistent best basis selection a la maxout network. In fact, our thresholding technique can be linked to an optimised ReLU based thresholding. In this scheme, applying best basis selection based on the empirical risk would thus become equivalent to the pooling operator of a maxout network.

REFERENCES

M Afifi, A Fassi-Fihri, M Marjane, K Nassim, M Sidki, and S Rachafi. Paul wavelet-based algorithm for optical phase distribution evaluation. *Optics communications*, 211(1):47–51, 2002.

Joakim Andén and Stéphane Mallat. Multiscale scattering for audio classification.

Joakim Andén, Vincent Lostanlen, and Stéphane Mallat. Joint time-frequency scattering for audio classification. In *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*, pp. 1–6. IEEE, 2015.

J-P Antoine, Pierre Carrette, Romain Murenzi, and Bernard Piette. Image analysis with two-dimensional continuous wavelet transform. *Signal processing*, 31(3):241–272, 1993.

Randall Balestriero and Behnaam Aazhang. Robust unsupervised transient detection with invariant representation based on the scattering network. *arXiv preprint arXiv:1611.07850*, 2016.

Randall Balestriero and Hervé Glotin. Scattering decomposition for massive signal classification: from theory to fast algorithm and implementation with validation on international bioacoustic benchmark. In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, pp. 753–761. IEEE, 2015.

Randall Balestriero and Herve Glotin. Linear time complexity deep fourier scattering network and extension to nonlinear invariants. *arXiv preprint arXiv:1707.05841*, 2017.

Kathrin Berkner. A correlation-dependent model for denoising via nonorthogonal wavelet transforms. 1998.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Joan Bruna and Stéphane Mallat. Classification with scattering operators. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1561–1566. IEEE, 2011.

Satinder Chopra* and Kurt J Marfurt. Choice of mother wavelets in cwt spectral decomposition. In *SEG Technical Program Expanded Abstracts 2015*, pp. 2957–2961. Society of Exploration Geophysicists, 2015.

Leon Cohen. *Time-frequency Analysis: Theory and Applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1995. ISBN 0-13-594532-1.

Ronald R Coifman and M Victor Wickerhauser. Entropy-based algorithms for best basis selection. *Information Theory, IEEE Transactions on*, 38(2):713–718, 1992.

R. Cosentino, R. Balestriero, and B. Aazhang. Best basis selection using sparsity driven multi-family wavelet transform. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 252–256, Dec 2016. doi: 10.1109/GlobalSIP.2016.7905842.

Ingrid Daubechies, Alex Grossmann, and Yves Meyer. Painless nonorthogonal expansions. *Journal of Mathematical Physics*, 27(5):1271–1283, 1986.

Paramveer S. Dhillon, Dean Foster, and Lyle H. Ungar. Minimum description length penalization for group and multi-task sparse learning. *J. Mach. Learn. Res.*, 12:525–564, February 2011. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1953048.1953064`.

David L Donoho, Iain M Johnstone, et al. Ideal denoising in an orthonormal basis chosen from a library of bases.

David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 301–369, 1995.

Costanza DAvanzoa, Vincenza Tarantinob, Patrizia Bisiacchib, and Giovanni Sparacinoa. A wavelet methodology for eeg time-frequency analysis in a time discrimination task.

Marie Farge. Wavelet transforms and their applications to turbulence. *Annual review of fluid mechanics*, 24(1):395–458, 1992.

James L Flanagan. Models for approximating basilar membrane displacement. *Bell Labs Technical Journal*, 39(5):1163–1191, 1960.

Pierre Goupillaud, Alex Grossmann, and Jean Morlet. Cycle-octave and related transforms in seismic signal analysis. *Geoexploration*, 23(1):85–102, 1984.

Alexander Grossmann and Jean Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM journal on mathematical analysis*, 15(4):723–736, 1984.

Hamid Krim, Dewey Tucker, Stephane Mallat, and David Donoho. On denoising and best signal representation. *IEEE transactions on information theory*, 45(7):2225–2238, 1999.

Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pensky. Sparse convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 806–814, 2015.

Vincent Lostanlen. *Opérateurs convolutionnels dans le plan temps-fréquence*. PhD thesis, Paris Sciences et Lettres, 2017.

Vincent Lostanlen and Joakim Andén. Binaural scene classification with wavelet scattering.

Vincent Lostanlen and Stéphane Mallat. Wavelet scattering on the pitch spiral. *arXiv preprint arXiv:1601.00287*, 2016.

Stéphane Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Stéphane Mallat. Understanding deep convolutional networks. *Phil. Trans. R. Soc. A*, 374(2065): 20150203, 2016.

Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

Sharan Narang, Gregory Diamos, Shubho Sengupta, and Erich Elsen. Exploring sparsity in recurrent neural networks. *arXiv preprint arXiv:1704.05119*, 2017.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

Markus Thom and Günther Palm. Sparse activity and sparse connectivity in supervised learning. *Journal of Machine Learning Research*, 14(Apr):1091–1143, 2013.

Christopher Torrence and Gilbert P Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1):61–78, 1998.

Arun Venkitaraman, Aniruddha Adiga, and Chandra Sekhar Seelamantula. Auditory-motivated gammatone wavelet transform. *Signal Processing*, 94:608–619, 2014.

Irene Waldspurger. Exponential decay of scattering coefficients. In *Sampling Theory and Applications (SampTA), 2017 International Conference on*, pp. 143–146. IEEE, 2017.

Dedy Rahman Wijaya, Riyanarto Sarno, and Enny Zulaika. Information quality ratio as a novel metric for mother wavelet selection. *Chemometrics and Intelligent Laboratory Systems*, 160: 59–71, 2017.

A Building a Deep Croisé Scattering Network

A.1 Continous Wavelet Transform

"By oscillating it resembles a wave, but by being localized it is a wavelet".

Yves Meyer

Wavelets were first introduced for high resolution seismology Goupillaud et al. (1984) Grossmann & Morlet (1984) and then developed theoretically by Meyer et al. Daubechies et al. (1986). Formally, wavelet is a function $\psi \in \mathbb{L}^2$ such that:

$$\int \psi(t)dt = 0, \tag{16}$$

it is normalized such that $\|\psi\|_{\mathbb{L}^2} = 1$. There exist two categories of wavelets, the discrete wavelets and the continuous ones. The discrete wavelets transform are constructed based on a system of linear equation. These equations represent the atom's property. These wavelet when scaled in a dyadic fashion form an orthonormal atom dictionary. Withal, the continuous wavelets have an explicit formulation and build an over-complete dictionary when successively scaled. In this work, we will focus on the continuous wavelets as they provide a more complete tool for analysis of signals. In order to perform a time-frequency transform of a signal, we first build a filter bank based on the mother wavelet. This wavelet is names the mother wavelet since it will be dilated and translated in order to create the filters that will constitute the filter bank. Notice that wavelets have a constant-Q property, thereby the ratio bandwidth to center frequency of the children wavelets are identical to the one of the mother. Then, the more the wavelet atom is high frequency the more it will be localized in time. The usual dilation parameters follows a geometric progression and belongs to the following set:

$$\Lambda = \left\{2^{j/Q}, j = 0, ..., J \times Q - 1\right\}$$

. Where the integers J and Q denote respectively the number of octaves, and the number of wavelets per octave. In order to develop a systematic and general principle to develop a filter bank for any wavelet family, we will consider the weighted version of the geometric progression mentioned above, that is:

$$\Lambda = \left\{\alpha 2^{j/Q}, j = 0, ..., J \times Q - 1\right\}$$

. In fact, the implementation of wavelet filter bank can be delicate since the mother wavelet has to be define at a proper center frequency such that no artifact or redundant information will appear in the final representation. Thus, in the section A.3 we propose a principled approach that allows the computation of the filter bank of any continuous wavelet. Beside, this re-normalized scaled is crucial to the comparison between different continuous wavelet. Having selected a geometric progression ensemble, the dilated version of the mother wavelet in the time are computed as follows:

$$\psi_\lambda(t) = \frac{1}{\lambda}\psi(\frac{t}{\lambda}),\ \forall \lambda \in \Lambda$$

, and can be calculated in the Fourier domain as follows:

$$\hat{\psi}_\lambda(\omega) = \hat{\psi}(\lambda\omega),\ \forall \lambda \in \Lambda$$

.

Notice that in practice the wavelets are computed in the Fourier domain as the wavelet transform will be based on a convolution operation which can be achieved with more efficiency. By construction the children wavelets have the same properties than the mother one. As a result, in the Fourier domain:

$$\hat{\psi}_\lambda = 0,\ \forall \lambda \in \Lambda$$

. Thus, to create a filter bank that cover all the frequency support, one needs a function that captures the low frequencies contents. The function is called the scaling function and satisfies the following criteria:

$$\int \phi(t)dt = 1$$

.

Finally, we denote by Wx, where $W \in \mathbb{C}^{N*(J*Q)\times N}$ is a block matrix such that each block corresponds to the filters at all scales for a given time. Also, we denote by $S(Wx)(\lambda, t)$ the reshape operator such that,

$$S(Wx)(\lambda, t) = (\frac{1}{\sqrt{\lambda}}\psi_\lambda^\star \star x)(t), \forall \lambda \in \Lambda, \tag{17}$$

where $\psi^\star$ is the complex conjugate of ψ_λ.

A.2 WAVELET FAMILIES

Among the continuous wavelets, different selection of mother wavelet is possible. Each one posses different properties, such as bandwidth, center frequency. This section is dedicated to the development of the families that are important for the analysis of diverse signals.

A.2.1 THE MORLET WAVELET

The Morlet wavelet (Fig. 6) is built by modulating a complex exponential and a Gaussian window defined in the time domain by,

$$\psi^M(t) = \pi^{-\frac{1}{4}} e^{i\omega_0 t} e^{-\frac{t^2}{2}}, \tag{18}$$

where ω_0 defines the frequency plane. In the frequency domain, we denote it by $\hat{\psi}^M(t)$,

$$\hat{\psi}^M(\omega) = \pi^{-\frac{1}{4}} e^{-\frac{(\omega-\omega_0)^2}{2}}, \forall \omega \in \mathbb{R}_+^\star, \tag{19}$$

thus, it is clear that ω_0 defines the center frequency of the mother wavelet.

With associated frequency center and standard deviation denoted respectively by $\omega_c^{\lambda_i}$ and $\Delta^{\lambda_i}\omega$, $\forall j \in \{0, ..., J * Q - 1\}$ are:

$$\omega_c^{\lambda_i} = \frac{\omega_0}{\lambda_i},$$
$$\Delta^{\lambda_i}\omega = \frac{1}{2\lambda_i^2}.$$

Notice that for the admissibility criteria $\omega_0 = 6$, however one can impose that zeros-mean condition facilely in the Fourier domain. Usually, this parameter is assign to the control of the center frequency of the mother wavelet, however in our case, we will see in the section A.3 a simple way to select a mother wavelet close enough to the Nyquist frequency such that all its contracted versions are properly defined. Then, we are able to vary the parameter ω_0 in order to have different support of Morlet wavelet.

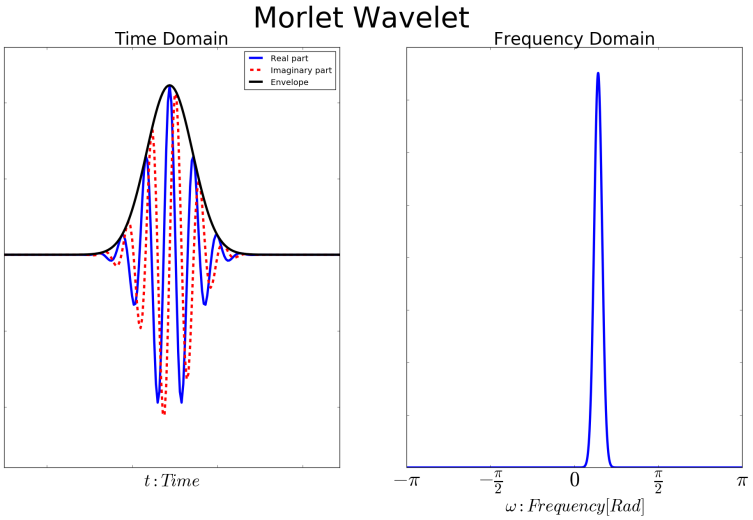


Figure 6: On the left a Morlet wavelet in the time domain where the dashed line is the imaginary part, the solid line is the real part, and the black envelope is the complex modulus, on the right a Morlet wavelet in the frequency domain.

The Morlet wavelet, is optimal from the uncertainty principle point of view Mallat (1999). The uncertainty principle, when given a time-frequency atoms, is the area of the rectangle of its joint time-frequency resolution. In the case of wavelet, given the fact that their ratio bandwidth to center frequency is equal implies that this area is equal for the mother wavelets and its scaled versions. As

a result, because of its time-frequency versatility this wavelet is wildly used for biological signals such as bio-acoustic Balestriero & Glotin (2015), seismic traces Chopra* & Marfurt (2015), EEG DAvanzoa et al. data.

A.2.2 THE GAMMATONE WAVELET

The Gammatone wavelet is a complex-valued wavelet that has been developed by Venkitaraman et al. (2014) via a transformation of the real-valued Gammatone auditory filter which provides a good approximation of the basilar membrane filter Flanagan (1960). Because of its origin and properties, this wavelet has been successfully applied for classification of acoustic scene Lostanlen & Andén. The Gammatone wavelet (Fig. 7) is defined in the time domain by,

$$\psi^G(t) = \left(2\pi(i-\sigma)t^{m-1} + (m-1)t^{m-2}\right) e^{-2pi\sigma t} e^{2piit}, \tag{20}$$

and in the frequency domain by,

$$\hat{\psi}^G(\omega) = \frac{i\omega(m-1)!}{(\sigma + i(\omega-\sigma))^m}. \tag{21}$$

A precise work on this wavelet achieved by V. Lostalnen in Lostanlen (2017) allows us to have an explicit formulation of the parameter σ such that the wavelet can be scaled while respecting the admissibility criteria:

$$\sigma^2 = \frac{r^{\frac{2}{m}}(1-r^{\frac{2}{m}})m^2\xi^2}{2}\left(\sqrt{1+\frac{B^2}{(1-r^{\frac{2}{m}})^2 m^2\xi^2}} - 1\right),$$

where ξ is the center frequency and B is the bandwidth parameter. Notice that $B = (1-2^{-\frac{1}{Q}})\xi$ with $\xi = \frac{2\pi}{1+2^{\frac{1}{Q}}}$ induce a quasi orthogonal filter bank. The associated frequency center and standard deviation denoted respectively by $\omega_c^{\lambda_i}$ and $\Delta^{\lambda_i}\omega$, $\forall j \in \{0, ..., J*Q-1\}$ are thus:

$$\omega_c^{\lambda_i} = \xi,$$
$$\Delta^{\lambda_i}\omega = B.$$

For this wavelet, thanks to the derivation in Lostanlen (2017), we can manually select for each order m the center frequency and bandwidth of the mother wavelet, which ease the filter bank design.

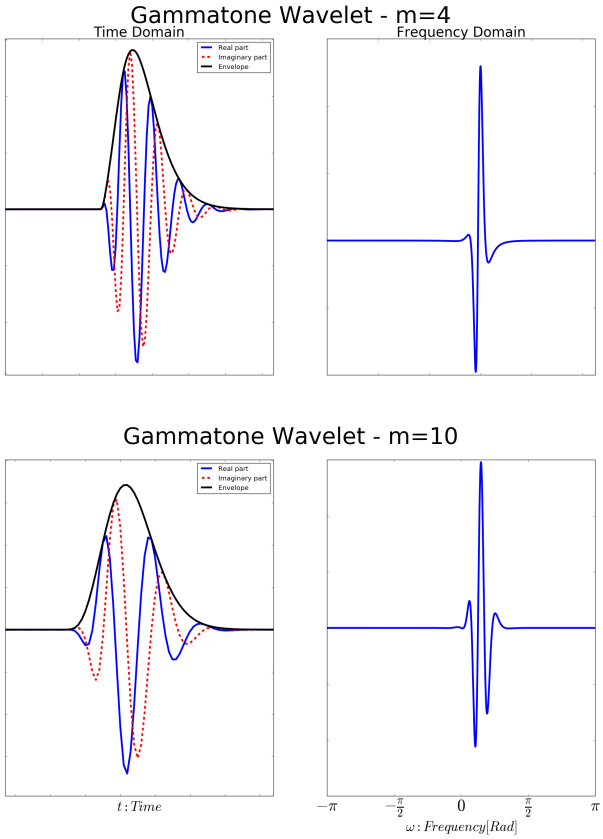


Figure 7: On the upper (bottom) left a $m = 4$ ($m = 10$) Gammatone wavelet in the time domain where the dashed line is the imaginary part and the solid line is the real part, on the upper (bottom) right a $m = 4$ ($m = 10$) wavelet in the frequency domain.

An important property that is directly related to the auditory response system is the asymmetric envelop, thereby the Gammatone wavelet is not invariant to time reversal to the contrary of the Morlet wavelet that behaves as a Gaussian function. Thus, for task such as sound classifications, this wavelet provides an efficient filter that will be prone to perceive the sound attack's. Beside this suitable property for specific analysis, this wavelet is near optimal with respect to the uncertainty principle. Notice that, when $m \to \infty$ it yields the Gabor wavelet Cohen (1995). Another interesting property of this wavelet is the causality, by taking into account only the previous and present information, there is no bias implied by some future information and thus it is suitable for real time signal analysis.

A.2.3 THE PAUL WAVELET

The Paul wavelet is a complex-valued wavelet which is highly localized in time, thereby has a poor frequency resolution. Because of its precision in the time domain, this wavelet is an ideal candidate to perform transient detection. The Paul wavelet of order m (Fig. 8) is defined in the time domain by,

$$\psi^P(t) = \frac{2^m i^m m!}{\sqrt{2m!\pi}}(1 - it)^{-(m+1)} \tag{22}$$

and in the frequency domain by,

$$\hat{\psi}^P(t) = \frac{2^m}{\sqrt{m(2m-1)!}}(\omega)^m e^{-\omega}, \forall \omega \in \mathbb{R}_+^\star, \tag{23}$$

With associated frequency center and standard deviation denoted respectively by $\omega_c^{\lambda_i}$ and $\Delta^{\lambda_i}\omega$, $\forall j \in \{0, ..., J * Q - 1\}$ are:

$$\omega_c^{\lambda_j} = \frac{2m+1}{2\lambda_j},$$
$$\Delta^{\lambda_j}\omega = \frac{\sqrt{2m+1}}{2\lambda_j}.$$

In Torrence & Compo (1998) they provide a clear and explicit formulation of some wavelet families applied the Paul wavelet in order to capture irregularly periodical variation in winds and sea surface temperatures over the tropical eastern Pacific Ocean . In addition, it directly represents the phase gradient from a single fringe pattern, yet providing a powerful tool in order to perform optical phase evaluation Afifi et al. (2002).

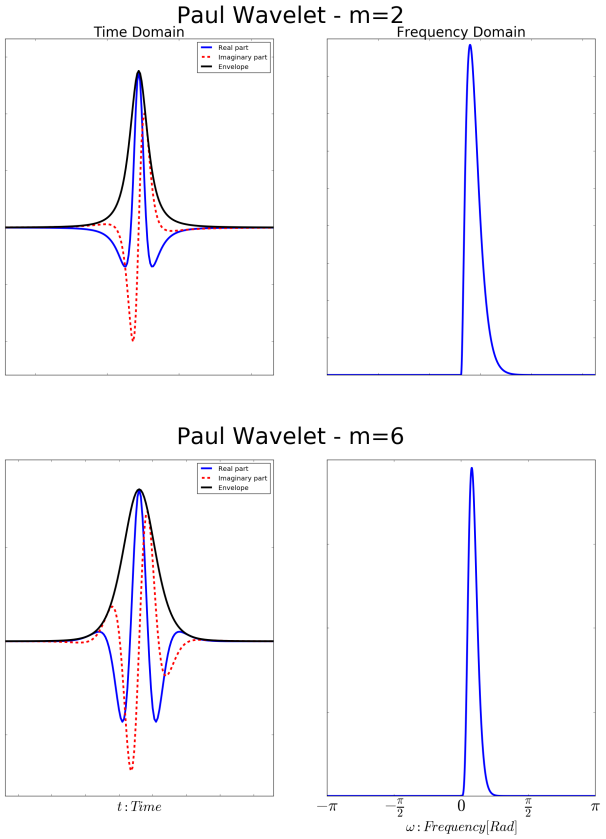


Figure 8: On the upper (bottom) left a $m = 2$ ($m = 6$) Paul wavelet in the time domain where the dashed line is the imaginary part and the solid line is the real part, on the upper (bottom) right a $m = 2$ ($m = 6$) wavelet in the frequency domain.

A.3 Filter Bank Design

In the previous section, we defined and develop the properties of several families of wavelets. Thereby, we can now consider the creation of the filter bank by means of these wavelets. Notice that we propose a simple manner to obtain the filter bank in the Fourier domain. Two reasons are at the origin of this choice: first, the wavelet transform is often computed in the Fourier domain because of its efficiency, secondly the wavelets are derived according to geometric progression scales, these scales can directly be represented in the frequency domain, thereby it provided us a way of knowing the position of the wavelet. However, in the time domain they are not directly quantifiable. Our systematic framework is based on the intuitive consideration of the problem: we have to select a

wavelet, named mother wavelet, that when contracted will create the filter bank derived from the selected scales. Assuming that the signals we will use are real valued, then the information represented in $[-\pi, 0]$ and $[0, \pi]$ are the same if extracted with a symmetric atom. Now, two kind of wavelets are considered, if the wavelet is complex-valued then its support is in $[0, \pi]$, thus the choice of the mother wavelet should be around π and the contracted all along the frequency axis until the total number of octave are covered. In the case of real-valued wavelet, if the wavelet is not symmetric then it will capture other phase information in the frequency band:$[-\pi, 0]$. Still, the mother wavelet can be selected to be close to π for its positive part, and $-\pi$ for its negative one. After defining the routine in order to select the mother wavelet, we propose a simple way to set the position of the mother wavelet. For each family, the center frequency and standard deviation are derived by finding α such that:

$$\omega_c^{\lambda_0} + \Delta^{\lambda_0}\omega = \pi, \tag{24}$$

where $\lambda_0 = \alpha * 2^{0/Q}$ denotes the first wavelet position. Given this equation, one create the mother wavelet such that it avoids capturing elements after the Nyquist frequency and avert the spectral mirror effect and artifacts. Given the value of α for a wavelet family, one can derive the wavelet filter bank according to the Algorithm 1. The wavelet filter banks generated by this algorithm for the different families aforementioned can be seen in Fig. 9. Notice that for sake of clarity, the scaling functions are not shown in Fig. 9.

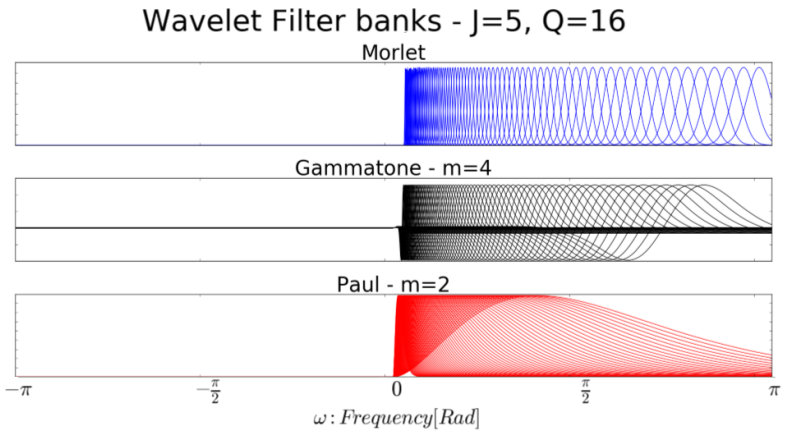


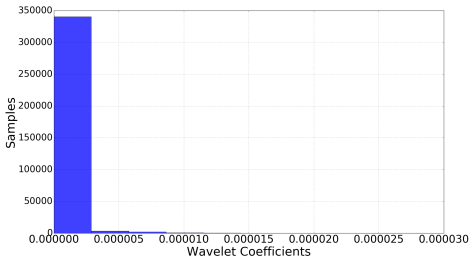
Figure 9: **From top to bottom**: Morlet wavelet filter bank, Gammatone wavelet filter bank with $m = 4$, Paul wavelet filter bank $m = 2$

Input: wavelet family: $f \in \mathcal{F}$, signal length: N, number of octaves: J, number of wavelets per octave: Q, scale weight: α, wavelet parameter: m
Output: D : Filter bank of the wavelet family f
Initialize the wavelet frequency domain: $\omega \in [-\pi, \pi]$
while $j < J * Q$ **do**
  $\lambda_j = \alpha^f 2^{\frac{j}{Q}}$ - Set up the scale for the jth children wavelet -
  $D_j = \psi^f(\lambda_j * \omega)$ - Compute the children wavelet at the given scale λ_j -
end

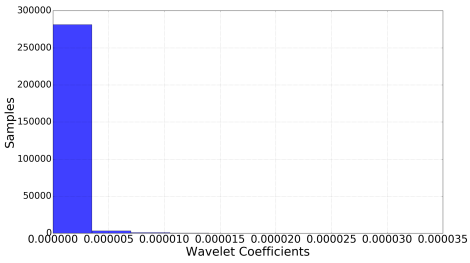
Algorithm 1: Compute Filter bank for any continuous wavelet family $f \in \mathcal{F}$

Finally, in order to guarantee the admissibility criterion one has to verify that all the wavelets are zeros-mean and square norm one. The first one is easily imposed by setting the wavelet to be null around $\omega = 0$ as it has been done to efficiently use the Morlet wavelet by Antoine et. al Farge (1992); Antoine et al. (1993). Then, because of Parseval equality and the energy conservation principle, the second one can be achieved by a re-normalization in the frequency domain of each atom.

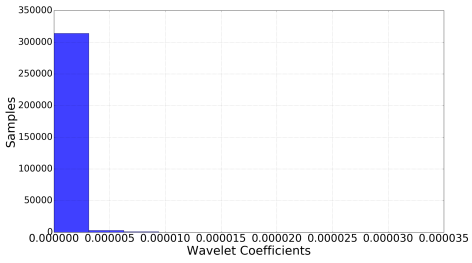
B ACTIVATION HISTOGRAM: SPARSITY EVALUATION LAYER 2



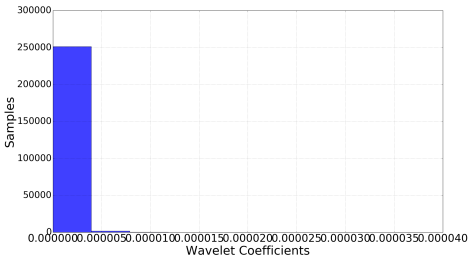
(a) SCSN 2nd Layer
Gammatone → Gammatone



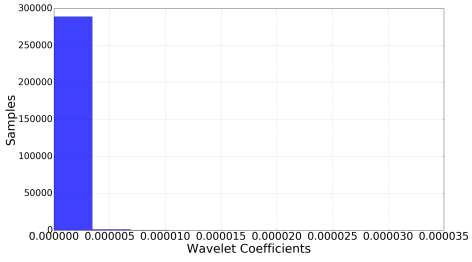
(b) CSN 2nd Layer
Gammatone → Gammatone



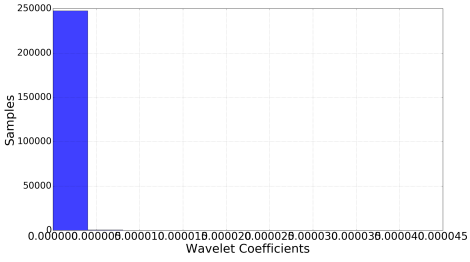
(c) SCSN 2nd layer
Gammatone → Morlet



(d) CSN 2nd Layer
Gammatone → Morlet



(e) SCSN 2nd Layer
Gammatone → Paul



(f) CSN 2nd Layer
Gammatone → Paul

Figure 10: **Histogram of activation of the second layer given in schema**

C DENOISING IN AN ORTHOGONAL BASIS FRAMEWORK

Assuming that the observed signal y, is corrupted with white noise,

$$y = x + \epsilon, \tag{25}$$

where ϵ is a vector of i.i.d centered normal distributions $\mathcal{N}(0, \sigma^2)$. Now we define the estimate of x by $\hat{x}_{W,D}$ such that:

$$\hat{x}_{W,D}(y) = W^T D^S W y \tag{26}$$

where W denotes the orthogonal basis and D^S is a diagonal binary operator such that,

$$D^S_{i,i} = \delta_i = \begin{cases} 1 \ if \ \ i \in S \\ 0 \ if \ \ i \in U \end{cases} \tag{27}$$

, where U and S denote respectively the set of selected and unselected wavelet coefficients. We also define D^U such that $I = D^U + D^S$. This estimate corresponds to a thresholding operation in the new basis and the inverse transform of this truncated representation.

We define the denoising problem as the solution of the following mean-square error:

$$\mathcal{R}^{\star}_{o}(x,W) = \min_{\delta} \mathbb{E}\left\|x - \hat{x}_{W,D}(y)\right\|^2 \tag{28}$$

$$= \min_{\delta} \mathbb{E}\left\|W^T(Wx - D^S W y\right\|^2 \tag{29}$$

$$= \min_{\delta} \mathbb{E}\left\|D^U Wx - D^S W(x+\epsilon)\right\|^2 \tag{30}$$

$$= \min_{\delta} \left\|D^U Wx\right\|^2 + \sigma^2 tr(D^S W W^T D^S) \tag{31}$$

$$= \min_{\delta} \sum_{i}^{n} [Wx]_i^2 1_{\{\delta_i=0\}} + \sigma^2 1_{\{\delta_i=1\}} \tag{32}$$

$$= \sum_{i}^{n} \min([Wx]_i^2, \sigma^2). \tag{33}$$

Therefore, the optimal $D^{S^\star}$ and $D^{U^\star}$ given by the following δ values:

$$\delta_i = 1_{|[Wx]_i^2| > \sigma^2}. \tag{34}$$

C.1 Upper-bound Non-orthogonal Risk & Empirical Risk

$$\hat{x}_{W,D}(y) = W^{\dagger} D^S W y \tag{35}$$

$$\mathcal{R}^{\star}(x,W) = \min_{\delta} \mathbb{E}\left\|x - \hat{x}_{W,D}(y)\right\|^2 \tag{36}$$

$$= \min_{\delta} \mathbb{E}\left\|W^{\dagger}(Wx - D^S W y\right\|^2 \tag{37}$$

$$= \min_{\delta} \mathbb{E}\left\|W^{\dagger}(D^U Wx - D^S W\epsilon)\right\|^2 \tag{38}$$

$$= \min_{\delta} \left\|W^{\star} D^U Wx\right\|^2 + \sigma^2 tr(W^T D^S W^{\dagger} W^{\dagger^T} D^S W), \tag{39}$$

Developing the previous expression and denoting by $\mu = Wx$ the wavelet coefficient vector, we have:

$$\begin{aligned}\mathcal{R}^{\star}(x,W) = &\min_{\delta} \sum_{t=1}^{n} \sum_{i,j=1}^{n(J*Q+1)} \mu_i \mu j \psi_t^{\dagger}[i] \psi_t^{\dagger}[j] 1_{\{\delta_i=0,\delta_j=0\}} \\ &+ \sigma^2 \sum_{i,j=1}^{n(J*Q+1)} (\sum_{t=1}^{n} \psi_t^{\dagger}[i]\psi_t^{\dagger}[j]) \psi_i^T \psi_j 1_{\{\delta_i=1,\delta_j=1\}}.\end{aligned} \tag{40}$$

we first use the triangular inequality,

$$\begin{aligned}\mathcal{R}^{\star}(x,W) \leq &\min_{\delta} \sum_{i,j=1}^{n(J*Q+1)} \left|\sum_{t=1}^{n} \mu_i \mu j \psi_t^{\dagger}[i] \psi_t^{\dagger}[j]\right| 1_{\{\delta_i=0,\delta_j=0\}} \\ &+ \sigma^2 \sum_{i,j=1}^{n(J*Q+1)} \left|\sum_{t=1}^{n} \psi_t^{\dagger}[i]\psi_t^{\dagger}[j]) \psi_i^T \psi_j\right| 1_{\{\delta_i=1,\delta_j=1\}}\end{aligned} \tag{41}$$

Now let's,

$$\mathcal{R}^U = \sum_{i,j=1}^{n(J*Q+1)} \left|\sum_{t=1}^{n} \mu_i \mu j \psi_t^{\dagger}[i] \psi_t^{\dagger}[j]\right|, \tag{42}$$

and,

$$\mathcal{R}^S = \sigma^2 \sum_{i,j=1}^{n(J*Q+1)} \left|\sum_{t=1}^{n} \psi_t^{\dagger}[i]\psi_t^{\dagger}[j]) \psi_i^T \psi_j\right|. \tag{43}$$

Then, based on the following min-max formulation, we obtain an upper bound of the ideal risk, that, when minimized will approximate the ideal risk in the overcomplete case:

$$\mathcal{R}^{\star}(x,W) \leq \sum_{k=1}^{n(J*Q+1)} \min_{\delta_k} \max_{\delta_l, l\neq k} \mathcal{R}^U 1_{\{\delta_i=0,\delta_j=0\}} + \mathcal{R}^S 1_{\{\delta_i=1,\delta_j=1\}} \tag{44}$$

$$\leq \sum_{k=1}^{n(J*Q+1)} \min_{\delta_k} (\max_{\delta_l, l\neq k} \mathcal{R}_1 + \max_{\delta_l, l\neq k} \mathcal{R}_2) \tag{45}$$

$$= \sum_{k=1}^{n(J*Q+1)} \min_{\delta_k} \; 1_{\{\delta_k=0\}} \left[\sum_{j=1}^{n*(J*Q+1)} \left| \sum_{t=1}^{n} \mu_k \mu j \psi_t^{\dagger}[k] \psi_t^{\dagger}[j] \right| \right]$$

$$+ 1_{\{\delta_k=1\}} \left[\sigma^2 \sum_{j=1}^{n(J*Q+1)} \left| \sum_{t=1}^{n} \psi_t^{\dagger}[k] \psi_t^{\dagger}[j]) \psi_k^T \psi_j \right| \right]. \tag{46}$$

Now, let's denote by $\mathcal{R}_{up}^U$ the error term corresponding to unselected coefficients:

$$\mathcal{R}_{up}^U = \sum_{j=1}^{n*(J*Q+1)} \left| \mu_k \mu j \sum_{t=1}^{n} \psi_t^{\dagger}[k] \psi_t^{\dagger}[j] \right|, \tag{47}$$

and by $\mathcal{R}_{up}^S$ for the selected ones:

$$\mathcal{R}_{up}^S = \sigma^2 \sum_{j=1}^{n(J*Q+1)} \left| \sum_{t=1}^{n} (\psi_t^{\dagger}[k] \psi_t^{\dagger}[j]) \psi_k^T \psi_j \right|. \tag{48}$$

we have that,

$$\mathcal{R}_{up}(x,W) = \sum_{k=1}^{n(J*Q+1)} \min_{\delta_k} \; 1_{\{\delta_k=0\}} \mathcal{R}_{up}^U + 1_{\{\delta_k=1\}} \mathcal{R}_{up}^S \tag{49}$$

$$= \sum_{k=1}^{n(J*Q+1)} \min(\mathcal{R}_{up}^U, \mathcal{R}_{up}^S). \tag{50}$$

C.2 Comparison Upper Bound Ideal Risk with Orthogonal Ideal Risk

Proposition 1.

Proof. The comparison of this upper bound risk given an orthogonal dictionary and the one derived in the orthogonal case is as follows:

If the basis is orthogonal, we have,

$$\sum_{t=1}^{n} (\psi_t^{\dagger}[k] \psi_t^{\dagger}[j]) = \begin{cases} 1, \; k=j \\ 0, \; else \end{cases} \tag{51}$$

and,

$$\psi_k^T \psi_j = \begin{cases} 1, \; k=j \\ 0, \; else \end{cases} \tag{52}$$

Therefore, the upper-bound derived recovers the ideal risk in the orthogonal case.

C.3 Comparison Upper Bound Ideal Risk with Empirical Risk

Proposition 2.

Proof. If $D^S = I$, the empirical risk is equal to:

$$\tilde{\mathcal{R}}(y, W) = \sigma^2 \sum_{j=1}^{n(J*Q+1)} \left| \sum_{t=1}^{n} (\psi_t^\dagger[k]\psi_t^\dagger[j])\psi_k^T \psi_j \right|.$$

and the upper bound risk is:

$$\mathcal{R}_{up}(x, W) = \sigma^2 \sum_{j=1}^{n(J*Q+1)} \left| \sum_{t=1}^{n} (\psi_t^\dagger[k]\psi_t^\dagger[j])\psi_k^T \psi_j \right|.$$

Thus both coincide as this restriction on the support of the risk makes it independent of both x and y.

Proposition 3.

Proof. In the case where $D^U = I$,

$$\tilde{\mathcal{R}}(y, W) = \sum_{j=1}^{n(J*Q+1)} \left| \mu_k(y)\mu_j(y) \sum_t \psi_t^\dagger[k]\psi_t^\dagger[j] \right| \tag{53}$$

$$= \sum_j |(\mu_k(x)\mu_j(x) + \mu_k(x)\mu_j(\epsilon) + \mu_k(\epsilon)\mu_k(x) + \mu_k(\epsilon)\mu_j(\epsilon))| \times \tag{54}$$

$$\left| \sum_t \psi_t^\dagger[k]\psi_t^\dagger[j] \right|, \tag{55}$$

by the triangular inequality, we have that:

$$\tilde{\mathcal{R}}(y, W) \leq \sum_j (|\mu_k(x)\mu_j(x)| + |\mu_k(x)\mu_j(\epsilon)| + |\mu_k(\epsilon)\mu_k(x)| + |\mu_k(\epsilon)\mu_j(\epsilon))| \times \tag{56}$$

$$\left| \sum_t \psi_t^\dagger[k]\psi_t^\dagger[j] \right|, \tag{57}$$

by the monotony of expectation and the Fubini theorem, we have almost surely:

$$\tilde{\mathcal{R}}(y, W) \leq \mathcal{R}_{up}(x, W) + C \times \left| \sum_{t=1}^{n} \psi_t^\dagger[k]\psi_t^\dagger[j] \right| \qquad a.s.,$$

where C is equals to,

$$C = \sum_{k=1}^{n(J*Q+1)} \sum_{j=1}^{n(J*Q+1)} |\mu_k(x)| \left\|\psi_j\right\|_1 \sigma\sqrt{\frac{2}{\pi}} + |\mu_j(x)| \left\|\psi_k\right\|_1 \sigma\sqrt{\frac{2}{\pi}} + \sigma^2(1 - \frac{2}{\pi}).$$