Instabilities in Convnets for Raw Audio

Daniel Haider, Vincent Lostanlen, Martin Ehler, and Peter Balazs

***Abstract*—What makes waveform-based deep learning so hard? Despite numerous attempts at training convolutional neural networks (convnets) for filterbank design, they often fail to outperform hand-crafted baselines. These baselines are linear time-invariant systems: as such, they can be approximated by convnets with wide receptive fields. Yet, in practice, gradient-based optimization leads to suboptimal approximations. In our article, we approach this phenomenon from the perspective of initialization. We present a theory of large deviations for the energy response of FIR filterbanks with random Gaussian weights. We find that deviations worsen for large filters and locally periodic input signals, which are both typical for audio signal processing applications. Numerical simulations align with our theory and suggest that the condition number of a convolutional layer follows a logarithmic scaling law between the number and length of the filters, which is reminiscent of discrete wavelet bases.**

***Index Terms*—Convolutional neural networks, digital filters, audio processing, statistical learning, frame theory.**

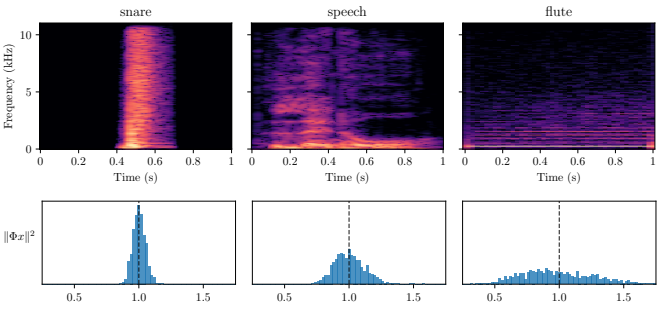

Fig. 1. Autocorrelation in the input signal $\boldsymbol{x}$ increases the variance of the filterbank response energy $\|\boldsymbol{\Phi x}\|^2$ across random initializations. We compare audio signals with different autocorrelation profiles. Left to right: Snare (low), speech (medium), and flute (high). Top: Spectrograms of the signals. Bottom: Empirical histogram of $\|\boldsymbol{\Phi x}\|^2$ for 1000 independent realizations of $\boldsymbol{\Phi}$.

I. Introduction

FILTERBANKS are linear time-invariant systems which decompose a signal $\boldsymbol{x}$ into $J > 1$ subbands. By convolution with filters $(\boldsymbol{w}_j)_{j=1,\dots,J}$ the output of a filterbank $\boldsymbol{\Phi}$ is a multivariate time series $(\boldsymbol{\Phi x})[n, j] = (\boldsymbol{x} * \boldsymbol{w}_j)[n]$. Filterbanks play a key role in speech and music processing: constant-Q-transforms, third-octave spectrograms, and Gammatone filterbanks are some well-known examples [1]–[3]. Beyond the case of audio, filterbanks are also used in other domains such as seismology [4], astrophysics [5], and neuroscience [6].

In deep learning, filterbanks serve as a preprocessing step to signal classification and generation. In this context, filterbank design is a form of feature engineering. Yet, in recent years, several authors have proposed to replace feature engineering with feature learning: i.e., to optimize filterbank parameters jointly with the rest of the pipeline [7]–[9].

So far, prior work on filterbank learning has led to mixed results. For example, on the TIMIT dataset, using a convolutional neural network (convnet) with 1-D filters on the "raw waveform" was found to fare poorly (29.2% phone error rate or PER) compared to the mel–spectrogram baseline (17.8% PER) [10]. Interestingly, fixing the convnet weights to form a filterbank on the mel–scale brings the PER to 18.3.%, and fine-tuning them by gradient descent, to 17.8%. Similar findings have been reported with Gammatone filterbanks [11].

Arguably, such a careful initialization procedure defeats the purpose of deep learning; i.e., sparing the effort of feature engineering. Furthermore, it contrasts with other domains (e.g., image processing) in which all layers may be initialized as random finite impulse responses (FIR). Yet, in audio processing, filterbank design may outperform filterbank learning, particularly from a random initialization; a fact that is increasingly well-documented [12]–[14]. A model known as multiresolution neural network (MuReNN) [15] has recently circumvented this issue in practice; however, the theory which underlies its empirical success remains unclear as of yet.

To understand the gap in performance in [10] and [11], we must distinguish neural network architecture design vs. iterative optimization. Simply put: just because a convnet *can* represent a human-engineered filterbank does not mean it *will*. This issue is not just of purely theoretical interest: in some emerging topics of machine listening such as bioacoustics, it would be practically useful to train a FIR filterbank with random initialization to learn something about acoustic events of interest with minimal domain-specific knowledge [16], [17].

Our article aims to explain the difficulties of deep learning in the raw waveform by offering a theoretical study of undecimated uniform filterbanks $\boldsymbol{\Phi}$ with large 1-D filters under random Gaussian initialization. Within the paradigm of filterbank learning, $\boldsymbol{\Phi}$ may be interpreted as the first layer of an untrained convnet with a stride of one. Prior publications have shown that stability is a crucial prerequisite for robustness to perturbations in the input [18] and stable dynamics in gradient-based optimization [19]. We characterize numerical stability in terms of energy preservation, i.e., when the ratio $(\|\boldsymbol{\Phi x}\|^2/\|\boldsymbol{x}\|^2)$ is close to one with high probability.

In Section II, we prove explicit formulas for the expected value and variance of $\|\boldsymbol{\Phi x}\|^2$, given a deterministic input sequence $\boldsymbol{x}$, and derive upper bounds for the probability of large deviations. In Section III, we bound the expected values and variances of the optimal frame bounds of $\boldsymbol{\Phi}$, i.e., $A = \min_{\|x\|=1} \|\boldsymbol{\Phi x}\|^2$ and $B = \max_{\|x\|=1} \|\boldsymbol{\Phi x}\|^2$. We conclude with an asymptotic analysis of the stability of $\boldsymbol{\Phi}$ by means of its condition number $\kappa = B/A$.

D. Haider and P. Balazs are with the Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria. V. Lostanlen is with Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France. M. Ehler is with University of Vienna, Faculty of Mathematics, Vienna, Austria.

II. FIR Filterbank with Random Gaussian Weights

Throughout this article, we use finite circulant convolution of signals $\boldsymbol{x} \in \mathbb{R}^N$ with filters $\boldsymbol{w} \in \mathbb{R}^T$, $T \leq N$, given by

$$(\boldsymbol{x} * \boldsymbol{w})[n] = \sum_{k=0}^{T-1} \boldsymbol{w}[k]\boldsymbol{x}[(n-k) \bmod N]. \quad (1)$$

We denote the circular autocorrelation of $\boldsymbol{x}$ for $0 \leq t < T$ by

$$\mathbf{R}_{\boldsymbol{xx}}(t) = \sum_{k=0}^{N-1} \boldsymbol{x}[k]\boldsymbol{x}[(k-t) \bmod N]. \quad (2)$$

A. Moments of the squared Euclidean norm

Proposition II.1. *Let $\boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{\Phi}$ a random filterbank with J i.i.d. filters $\boldsymbol{w}_j \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ of length $T \leq N$. Then*

$$\mathbb{E}\left[\|\boldsymbol{\Phi}\boldsymbol{x}\|^2\right] = JT\sigma^2\|\boldsymbol{x}\|^2 \quad (3)$$

and

$$\mathbb{V}\left[\|\boldsymbol{\Phi}\boldsymbol{x}\|^2\right] = 2J\sigma^4 \sum_{\tau=-T}^{T} \big(T - |\tau|\big)\mathbf{R}_{\boldsymbol{xx}}(\tau)^2. \quad (4)$$

We note that (3) is known for $J = 1$ and $T = N$ [20]. Setting $\sigma^2 = (JT)^{-1}$ implies $\mathbb{E}\left[\|\boldsymbol{\Phi}\boldsymbol{x}\|^2\right] = \|\boldsymbol{x}\|^2$. In other words, if the variance of each parameter $\mathbf{w}_j$ scales in inverse proportion with the total number of parameters (i.e., JT), then $\boldsymbol{\Phi}$ satisfies energy preservation on average. However, it is important to see that the variance of the random variable $\|\boldsymbol{\Phi}\boldsymbol{x}\|^2$ depends also on the content of the input $\boldsymbol{x}$: specifically, its autocorrelation $\mathbf{R}_{\boldsymbol{xx}}$. This is a peculiar property of convnets, unlike fully connected layers with random Gaussian initialization, see Proposition V.1 in the appendix. This can be explained by the fact that the entries of the random matrix associated with $\boldsymbol{\Phi}$ are not independent. In this context, we note that natural audio signals are often locally periodic and thus highly autocorrelated. Hence, we interpret Proposition II.1 as follows: untrained convnets are particularly unstable in the presence of vowels in speech or pitched notes in music. Figure 1 illustrates this phenomenon for three real-world signals. Our proof of Proposition II.1 hinges on the following lemmata, which are proven in the appendix.

Lemma II.2. *Let $\boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{w} \in \mathbb{R}^T$, $T \leq N$. The circular convolution of $\boldsymbol{x}$ and $\boldsymbol{w}$ satisfies $\|\boldsymbol{x} * \boldsymbol{w}\|^2 = \boldsymbol{w}^\top \mathbf{Q}_T(\boldsymbol{x})\boldsymbol{w}$, where the entries of the matrix $\mathbf{Q}_T(\boldsymbol{x})$ are given by $\mathbf{Q}_T(\boldsymbol{x})[n,t] = \mathbf{R}_{\boldsymbol{xx}}((t-n) \bmod N)$ for each $0 \leq n,t < T$.*

Lemma II.3. *Let $\boldsymbol{x} \in \mathbb{R}^N$. All diagonal entries of the matrix $\mathbf{Q}_T(\boldsymbol{x})$ from Lemma II.2 are equal to $\|\boldsymbol{x}\|^2$.*

Proof of Proposition II.1. Given a filter $\boldsymbol{w}_j$ for $1 \leq j \leq J$, we apply Lemma II.2 and use the cyclic property of the trace

$$\|\boldsymbol{x} * \boldsymbol{w}_j\|^2 = \mathrm{Tr}\left(\boldsymbol{w}_j^\top \mathbf{Q}_T(\boldsymbol{x})\boldsymbol{w}_j\right) = \mathrm{Tr}\left(\mathbf{Q}_T(\boldsymbol{x})\boldsymbol{w}_j\boldsymbol{w}_j^\top\right). \quad (5)$$

We take the expected value on both sides and recognize the term $\mathbb{E}[\boldsymbol{w}_j\boldsymbol{w}_j^\top]$ as the covariance matrix of $\boldsymbol{w}_j$, i.e., $\sigma^2\mathbf{I}$. Hence:

$$\mathbb{E}\left[\|\boldsymbol{x} * \boldsymbol{w}_j\|^2\right] = \mathrm{Tr}\left(\mathbf{Q}_T(\boldsymbol{x})\mathbb{E}\left[\boldsymbol{w}_j\boldsymbol{w}_j^\top\right]\right) = \sigma^2\,\mathrm{Tr}\left(\mathbf{Q}_T(\boldsymbol{x})\right). \quad (6)$$

By Lemma II.3, $\mathrm{Tr}(\mathbf{Q}_T(\boldsymbol{x})) = T\|\boldsymbol{x}\|^2$, hence $\mathbb{E}[\|\boldsymbol{x} * \boldsymbol{w}_j\|^2] = \sigma^2 T\|\boldsymbol{x}\|^2$. For the variance, we recall Theorem 5.2 from [21], which states that if $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for any matrix $\mathbf{A}$

$$\mathbb{V}\left[\boldsymbol{v}^\top\mathbf{A}\boldsymbol{v}\right] = 2\,\mathrm{Tr}\left(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}\right) + 4\boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu} \quad (7)$$

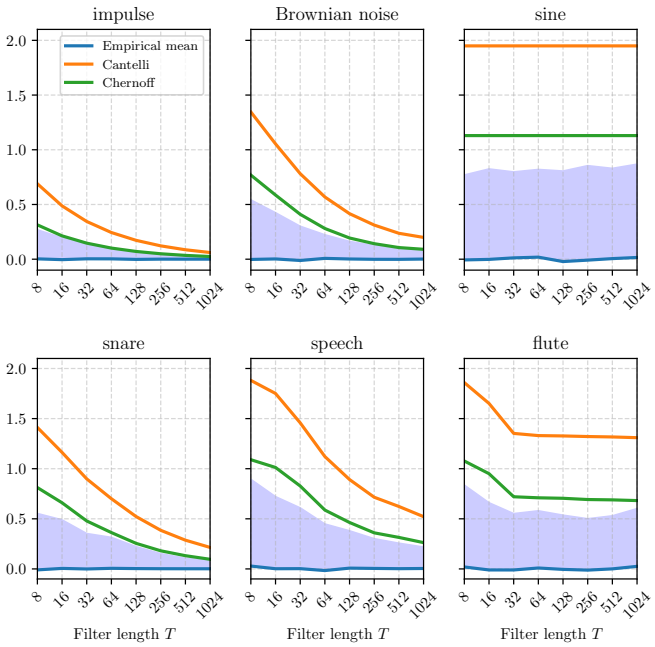

Fig. 2. Large deviations of filterbank response energy ($\|\boldsymbol{\Phi}\boldsymbol{x}\|^2 - \|\boldsymbol{x}\|^2$) for three synthetic signals of length $N = 1024$ (top) and three natural signals of length $N = 22050$ (bottom). Blue: empirical mean and 95$^{\text{th}}$ percentile across 1000 realizations of $\boldsymbol{\Phi}$. We show two theoretical bounds from Proposition II.4: Cantelli (Equation (9), orange) and Chernoff (Equation 10, green). Each filterbank contains $J = 10$ filters of length $T = 2^k$ where $3 \leq k \leq 10$.

We set $\boldsymbol{v} = \boldsymbol{w}_j$, $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, and $\mathbf{A} = \mathbf{Q}_T(\boldsymbol{x})$. We obtain:

$$\begin{aligned}\mathbb{V}\left[\|\boldsymbol{x} * \boldsymbol{w}_j\|^2\right] &= 2\sigma^4\,\mathrm{Tr}\left(\mathbf{Q}_T(\boldsymbol{x})^2\right) \\ &= 2\sigma^4 \sum_{t=0}^{T-1}\sum_{t'=0}^{T-1} \mathbf{R}_{\boldsymbol{xx}}(t'-t)\mathbf{R}_{\boldsymbol{xx}}(t-t') \\ &= 2\sigma^4 \sum_{t=0}^{T-1}\sum_{\tau=-t}^{T-1-t} \mathbf{R}_{\boldsymbol{xx}}(\tau)^2. \end{aligned} \quad (8)$$

By a combinatorial argument, the double sum above rewrites as $\sum_{\tau=-T}^{T}\big(T-|\tau|\big)\mathbf{R}_{\boldsymbol{xx}}(\tau)^2$. The proof concludes by linearity of the variance, given the independence of the J filters in $\boldsymbol{\Phi}$. ■

After scaling $\boldsymbol{\Phi}$ such that it preserves energy on average, i.e. $\mathbb{E}\left[\|\boldsymbol{\Phi}\boldsymbol{x}\|^2\right] = \|\boldsymbol{x}\|^2$, we now derive upper bounds on the probability of large deviations of $\|\boldsymbol{\Phi}\boldsymbol{x}\|^2$ given $\boldsymbol{x} \neq \mathbf{0}$.

Proposition II.4 (Cantelli bound)**.** *Let $\boldsymbol{\Phi}$ be a random filterbank with J i.i.d. filters $\boldsymbol{w}_j \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ of length T and $\sigma^2 = (JT)^{-1}$. Given a deviation $\alpha \geq 0$, the probability of $\|\boldsymbol{\Phi}\boldsymbol{x}\|^2$ exceeding $(1+\alpha)\|\boldsymbol{x}\|^2$ is bounded from above as*

$$\mathbb{P}\left[\|\boldsymbol{\Phi}\boldsymbol{x}\|^2 \geq (1+\alpha)\|\boldsymbol{x}\|^2\right] \leq \frac{\mathbb{V}\left[\|\boldsymbol{\Phi}\boldsymbol{x}\|^2\right]}{\mathbb{V}\left[\|\boldsymbol{\Phi}\boldsymbol{x}\|^2\right] + \alpha^2\mathbf{R}_{\boldsymbol{xx}}(0)^2}. \quad (9)$$

Proposition II.5 (Chernoff bound)**.** *Let $\boldsymbol{\lambda}$ denote the vector of eigenvalues of $\mathbf{Q}_T(\boldsymbol{x})$. Under the same assumptions as Proposition II.4, and given a deviation $\alpha \geq 0$, the probability of $\|\boldsymbol{\Phi}\boldsymbol{x}\|^2$ exceeding $(1+\alpha)\|\boldsymbol{x}\|^2$ is bounded from above as*

$$\mathbb{P}\left[\|\boldsymbol{\Phi}\boldsymbol{x}\|^2 \geq (1+\alpha)\|\boldsymbol{x}\|^2\right] \leq \exp\left(-\frac{\alpha^2 JT^2\|\boldsymbol{x}\|^4}{2\alpha T\|\boldsymbol{\lambda}\|_\infty\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{\lambda}\|_2^2}\right). \quad (10)$$

The two propositions above have their own merits. Proposition II.4, based on Cantelli's inequality [22], is straightforward and interpretable in terms of the autocorrelation of $\boldsymbol{x}$. Meanwhile, Proposition II.5, based on Chernoff's inequality [23], is closer to empirical percentiles, yet is expressed in terms of the eigenvalues of $\mathbf{Q}_T(\boldsymbol{x})$, for which there is no general formula. In the particular case of full-length filters ($T = N$), $\mathbf{Q}_T(\boldsymbol{x})$ is a circulant matrix: hence, we interpret these eigenvalues as the energy spectral density of the input signal, i.e., $\boldsymbol{\lambda} = |\hat{\boldsymbol{x}}|^2$ where $\hat{\boldsymbol{x}}$ is the discrete Fourier transform of $\boldsymbol{x}$.

B. Numerical simulation

We now compute empirical probabilities of relative energy deviations between $\|\boldsymbol{\Phi x}\|^2$ and $\|\boldsymbol{x}\|^2$ for different signals $\boldsymbol{x}$ and various filter lengths T. Specifically, for each $\boldsymbol{x}$ and each T, we simulate 1000 independent realizations of $\|\boldsymbol{\Phi x}\|^2$ for each value of T and retain the closest 95% displayed as shaded area in Figure 2. Additionally, we set the right-hand side of Propositions II.4 and II.5 to 5% and solve for α, yielding upper bounds for this area.

The upper part of Figure 2 illustrates our findings for three synthetic signals: (i) a single impulse, which has low autocorrelation, (ii) a realization of Brownian noise, which has medium autocorrelation and (iii) a sine wave with frequency $\omega = \pi$, which has high autocorrelation. In the lower part of the same figure, we use real-world sounds: a snare drum hit, a spoken utterance, and a sustained note on the concert flute.

As predicted by the theory, large deviations of $\|\boldsymbol{\Phi x}\|^2$ become less probable as the filters grow in length T if the input $\boldsymbol{x}$ has little autocorrelation (e.g., snare). The rate of decay is slower for highly autocorrelated signals (e.g., flute). These findings explain the observations we already made in Figure 1.

III. Extreme Value Theory meets Frame Theory

In the previous section, we have described the probability distribution of $\|\boldsymbol{\Phi x}\|^2$ for a known input signal $\boldsymbol{x}$. We now turn to inquire about the properties of $\boldsymbol{\Phi}$ as a linear operator; i.e., independently of $\boldsymbol{x}$. If there exist two positive numbers A and B such that the double inequality $A\|\boldsymbol{x}\|^2 \leq \|\boldsymbol{\Phi x}\|^2 \leq B\|\boldsymbol{x}\|^2$ holds for any $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{\Phi}$ is said to be a *frame* for $\mathbb{R}^N$ with frame bounds A and B. The optimal frame bounds are given by $A = \min_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{\Phi x}\|^2$ and $B = \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{\Phi x}\|^2$.

A. From quadratic forms to chi-squared distributions

Although the expected frame bounds $\mathbb{E}[A]$ and $\mathbb{E}[B]$ do not have closed-form expressions, we can relate them to the expected order statistics of the chi-squared distribution with J degrees of freedom, denoted by $\chi^2(J)$.

Theorem III.1. *Let $\boldsymbol{\Phi}$ be a random filterbank with J i.i.d. filters $\boldsymbol{w}_j \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ with $\sigma^2 = (JT)^{-1}$. The expectations of the optimal frame bounds A, B of $\boldsymbol{\Phi}$ are bounded by the order statistics of $Y_0, \ldots, Y_{T-1} \sim \chi^2(J)$ i.i.d., as follows*

$$J^{-1}\mathbb{E}[Y_T^{\min}] \leq \mathbb{E}[A] \leq 1 \leq \mathbb{E}[B] \leq J^{-1}\mathbb{E}[Y_T^{\max}], \qquad (11)$$

where $Y_T^{\min} = \min_{0\leq k<T} Y_k$ and $Y_T^{\max} = \max_{0\leq k<T} Y_k$.

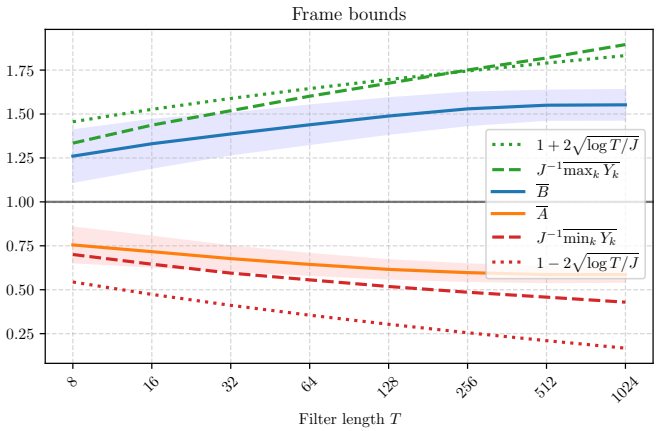

Fig. 3. Empirical means $\overline{A}$ and $\overline{B}$ (solid lines) and 95$^{\text{th}}$ percentiles (shaded area) of frame bounds A and B for 1000 instances of $\boldsymbol{\Phi}$ with $\sigma^2 = (TJ)^{-1}$, $J = 40$ and different values of T. Dashed lines denote the bounds of $\mathbb{E}[A]$ and $\mathbb{E}[B]$ from Theorem III.1. Dotted lines denote the asymptotic bounds proposed in (20).

Proof. The inner inequalities ($\mathbb{E}[A] \leq 1 \leq \mathbb{E}[B]$) are a direct consequence of Proposition II.1. Regarding the outer inequalities, we perform an eigenvalue decomposition of $\mathbf{Q}_T(\boldsymbol{x}) = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$, where the columns of $\mathbf{U}$ contain the eigenvectors of $\mathbf{Q}_T(\boldsymbol{x})$ as columns and the diagonal matrix $\boldsymbol{\Lambda}$ contains the spectrum of eigenvalues, $\boldsymbol{\lambda}$. For each filter $\boldsymbol{w}_j$ with $1 \leq j \leq J$, let us use the shorthand $\boldsymbol{y}_j = \mathbf{U}^\top \boldsymbol{w}_j$. By Lemma II.2 we obtain

$$\|\boldsymbol{x} * \boldsymbol{w}_j\|^2 = \boldsymbol{w}_j^\top \mathbf{U}^\top \boldsymbol{\Lambda} \mathbf{U} \boldsymbol{w}_j = \sum_{k=0}^{T-1} \lambda_k \boldsymbol{y}_j[k]^2. \qquad (12)$$

We define $Y_k = \sum_{j=1}^{J} (\boldsymbol{y}_j^2[k]/\sigma^2)$. Equation (12) yields

$$\|\boldsymbol{\Phi x}\|^2 = \sigma^2 \sum_{k=0}^{N-1} \lambda_k \sum_{j=1}^{J} \frac{\boldsymbol{y}_{j,k}^2}{\sigma^2} = \sigma^2 \sum_{k=0}^{N-1} \lambda_k Y_k. \qquad (13)$$

Since $\mathbf{Q}_T(\boldsymbol{x})$ is a real symmetric matrix, $\mathbf{U}$ is an orthogonal matrix. Thus, $\boldsymbol{y}_j$ follows the same distribution as $\boldsymbol{w}_j$

$$\mathbf{U}^\top \boldsymbol{w}_j \sim \mathcal{N}(0, \sigma^2 \mathbf{U}\mathbf{I}\mathbf{U}^\top) = \mathcal{N}(0, \sigma^2 \mathbf{I}). \qquad (14)$$

For all k with $0 \leq k < T$, $(\boldsymbol{y}_j^2[k]/\sigma^2)$ are i.i.d. standard Gaussian random variables. Thus, the Y_k's are also i.i.d. and follow a $\chi^2(J)$ distribution. Let us define the associated order statistics

$$Y_T^{\min} = \min_{0\leq k<T} Y_k \quad \text{and} \quad Y_T^{\max} = \max_{0\leq k<T} Y_k. \qquad (15)$$

Lemma II.3 implies $\sum_{k=0}^{T-1} \lambda_k = \mathrm{Tr}(\mathbf{Q}_T(\boldsymbol{x})) = T\|\boldsymbol{x}\|^2$. Hence

$$\begin{aligned} \min_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{\Phi x}\|^2 - \sigma^2 T Y_T^{\min} \geq 0, \\ \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{\Phi x}\|^2 - \sigma^2 T Y_T^{\max} \leq 0, \end{aligned} \qquad (16)$$

where the inequalities are understood as almost sure. Taking the expectation and setting $\sigma^2 = (JT)^{-1}$ yields the claim. ∎

In Figure 3, we have performed numerical simulations that align with the statement of Theorem III.1. We observe that optimal frame bounds A and B typically diverge away from 1 as T up grows to 2^{10}, a common value in audio applications.

This phenomenon is evidence of instabilities at initialization of a convnet and, consequently, also during training.

After bounding the expected values of A and B, we now turn to their variances. We refer to the appendix for a proof.

Proposition III.2. *Let $\mathbf{\Phi}$ be a random filterbank with J i.i.d. filters $\boldsymbol{w}_j \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ with $\sigma^2 = (JT)^{-1}$. The variances of the optimal frame bounds A and B can be bounded as*

$$2(TJ)^{-1} \leq \mathbb{V}[A], \mathbb{V}[B] \leq 2J^{-1}. \tag{17}$$

B. Asymptotics of the condition number

The ratio $\kappa = B/A$, known as condition number, characterizes the numerical stability of $\mathbf{\Phi}$. In particular, κ equals one if and only if there exists $C > 0$ such that $\|\mathbf{\Phi}\boldsymbol{x}\|^2 = C\|\boldsymbol{x}\|^2$. However, its expected value, $\mathbb{E}[\kappa]$, may be strictly greater than one even so $\mathbb{E}\left[\|\mathbf{\Phi}\boldsymbol{x}\|^2\right] = C\|\boldsymbol{x}\|^2$ holds for every $\boldsymbol{x}$. Since A and B are dependent random variables, $\mathbb{E}[\kappa]$ is difficult to study analytically [24]. We conjecture that $1 \leq \mathbb{E}[\kappa] \leq (\mathbb{E}[B]/\mathbb{E}[A])$, which is equivalent to $\mathrm{cov}(\kappa, A) \geq 0$.

Unfortunately, the expected values of $Y_T^{\min}$ and $Y_T^{\max}$ that are used for the bounds of $\mathbb{E}[A]$ and $\mathbb{E}[B]$ in Theorem III.1 are not available in closed form for finite values of T [25]. Nevertheless, for a large number of degrees of freedom J, $\chi^2(J)$ resembles a normal distribution with mean J and variance $2J$, such that we propose to replace $Y_T^{\min}$ and $Y_T^{\max}$ by

$$\tilde{Y}_T^{\min} = \min_{0 \leq k < T} \tilde{Y}_k \quad \text{and} \quad \tilde{Y}_T^{\max} = \max_{0 \leq k < T} \tilde{Y}_k, \tag{18}$$

where the $\tilde{Y}_k$'s are i.i.d. drawn from $\mathcal{N}(J, 2J)$ [26]. From the extreme value theorem for the standard normal distribution (see e.g. Theorem 1.5.3. in [27]) we know that for large T, we can asymptotically approximate the expectations of (18) by

$$\mathbb{E}\left[\tilde{Y}_T^{\min}\right] \propto J - 2\sqrt{J \log T} \quad \text{and} \quad \mathbb{E}\left[\tilde{Y}_T^{\max}\right] \propto J + 2\sqrt{J \log T}. \tag{19}$$

The equations above suggest approximate bounds for $\mathbb{E}[A]$ and $\mathbb{E}[B]$. We draw inspiration from them to propose the value

$$\tilde{\kappa}(J, T) = \left(1 + 2\sqrt{\frac{\log T}{J}}\right) \Bigg/ \left(1 - 2\sqrt{\frac{\log T}{J}}\right), \tag{20}$$

as asymptotic error bound for $\mathbb{E}[\kappa]$, subject to $T \to \infty$ and $J > 4 \log T$. Interestingly, the level sets of $\tilde{\kappa}$ satisfy $J \propto \log T$, a scaling law which is reminiscent of the theory underlying the construction of discrete wavelet bases [28].

C. Numerical simulation

Figure 4 (top) shows empirical means of κ for 1000 independent realizations of $\mathbf{\Phi}$ and various settings of J and T. Qualitatively speaking we observe that convnets with few long filters (small J, large T) suffer from ill-conditioning, as measured by a large κ. This is despite having set $\sigma^2 = (JT)^{-1}$, which implies that $\mathbf{\Phi}$ satisfies energy preservation on average (Proposition II.1). Figure 4 (bottom) shows the result of the same simulation with J on the horizontal axis, together with our proposed scaling law $J \propto \log T$. We observe that filterbanks that follow this scaling law have approximately the same condition number κ on average.

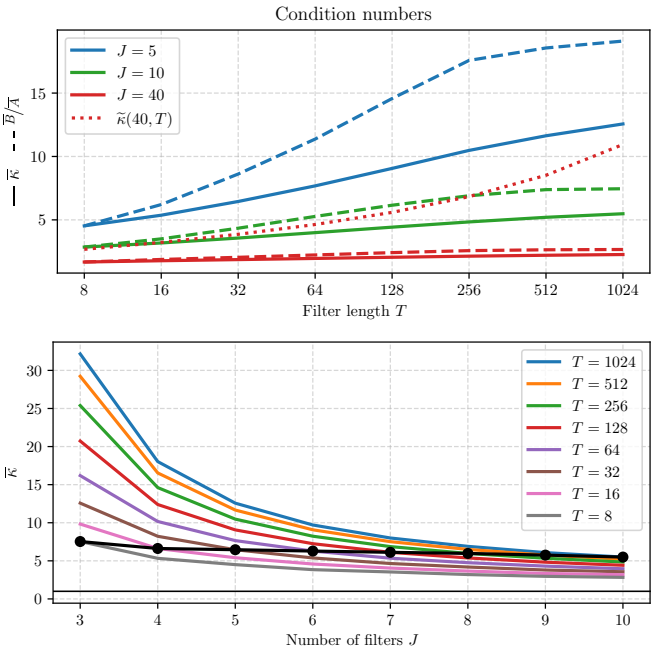

Fig. 4. We denote by $\overline{A}, \overline{B}$, and $\overline{\kappa}$ the empirical means of the respective quantities over 1000 instances of $\mathbf{\Phi}$ with $\sigma^2 = (TJ)^{-1}$. Top: Comparison of $\overline{\kappa}$ (solid) and $\overline{B}/\overline{A}$ (dashed) for increasing filter length T and different values of J. Bottom: Empirical mean $\overline{\kappa}$ for increasing numbers of filters J and different values T. For $J = \log_2 T$ (solid black), $\overline{\kappa}$ remains approximately constant.

IV. Conclusion

This article presents a large deviations theory of energy dissipation in random filterbanks. We have found that the variance of output energy $\|\mathbf{\Phi}\boldsymbol{x}\|^2$ grows with the autocorrelation of the input sequence $\boldsymbol{x}$ (Proposition II.1). Thus, natural audio signals, which typically have high short-term autocorrelation, are *adversarial examples* to 1-D convnets, in the sense that they trigger numerical instabilities with high probability. Furthermore, we have shown that numerical stability depends strongly on architecture design for $\mathbf{\Phi}$; specifically, the number of filters J and their lengths T. By combining frame theory with extreme value theory, we have explained why the most stable convnets are those with many short filters (large T, short J). For large convnets, we have identified a scaling law ($J \propto \log T$) which roughly preserves the condition number of $\mathbf{\Phi}$. Characterizing the probability distribution of the condition number for non-asymptotic values of J and T remains an open problem. As the next step, we plan to study the potential numerical instabilities that arise due to aliasing effects from strided convolution in decimated random filterbanks.[1]

Acknowledgment

D. Haider is recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Acoustics Research Institute (A 26355). V. Lostanlen is supported by ANR MuReNN. The work of M. Ehler was supported by the WWTF project CHARMED (VRG12-009) and P. Balazs was supported by the FWF projects LoFT (P 34624) and NoMASP (P 34922).

[1]The source code for reproducing all numerical simulations may be found at https://github.com/danedane-haider/Random-Filterbanks.

REFERENCES

[1] T. Necciari, N. Holighaus, P. Balazs, Z. Průša, P. Majdak, and O. Derrien, "Audlet filter banks: A versatile analysis/synthesis framework using auditory frequency scales," *Applied Sciences*, vol. 8, no. 1, 2018.

[2] P. Balazs, N. Holighaus, T. Necciari, and D. Stoeva, *Frame Theory for Signal Processing in Psychoacoustics*. Springer International Publishing, 2017, pp. 225–268.

[3] R. F. Lyon, *Human and machine hearing: Extracting meaning from sound*. Cambridge University Press, 2017.

[4] M.-A. Meier, T. Heaton, and J. Clinton, "The Gutenberg algorithm: Evolutionary Bayesian magnitude estimates for earthquake early warning with a filter bank," *Bulletin of the Seismological Society of America*, vol. 105, no. 5, pp. 2774–2786, 2015.

[5] E. Chassande-Mottin, "Learning approach to the detection of gravitational wave transients," *Physical Review D*, vol. 67, no. 10, 2003.

[6] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *Proc. IEEE IJCNN*, 2008, pp. 2390–2397.

[7] M. Dörfler, T. Grill, R. Bammer, and A. Flexer, "Basic filters for convolutional neural networks applied to music: Training or design?" *Neural Computing and Applications*, vol. 32, pp. 941–954, 2020.

[8] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE SLT*, 2018, pp. 1021–1028.

[9] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," in *Proc. ICLR*, 2021.

[10] N. Zeghidour, N. Usunier, I. Kokkinos, T. Schaiz, G. Synnaeve, and E. Dupoux, "Learning filterbanks from raw speech for phone recognition," in *Proc. IEEE ICASSP*, 2018, pp. 5509–5513.

[11] T. Sainath, R. J. Weiss, K. Wilson, A. W. Senior, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Proc. INTERSPEECH*, 2015.

[12] I. López-Espejo, Z.-H. Tan, and J. Jensen, "Exploring filterbank learning for keyword spotting," in *Proc. EUSIPCO*, 2021, pp. 331–335.

[13] J. Schlüter and G. Gutenbrunner, "EfficientLEAF: A faster learnable audio frontend of questionable use," in *Proc. EUSIPCO*. IEEE, 2022, pp. 205–208.

[14] F. J. Bravo Sanchez, M. R. Hossain, N. B. English, and S. T. Moore, "Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.

[15] V. Lostanlen, D. Haider, H. Han, M. Lagrange, P. Balazs, and M. Ehler, "Fitting auditory filterbanks with multiresolution neural networks," in *Proc. IEEE WASPAA*, 2023, pp. 1–5.

[16] S. L. Hopp, M. J. Owren, and C. S. Evans, *Animal acoustic communication: sound analysis and research methods*. Springer Science & Business Media, 2012.

[17] D. Stowell, "Computational bioacoustics with deep learning: A review and roadmap," *PeerJ*, vol. 10, 2022.

[18] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," *Proc. ICML*, 2017.

[19] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, http://www.deeplearningbook.org.

[20] M. Ehler, "Preconditioning filter bank decomposition using structured normalized tight frames," *Journal of Applied Mathematics*, vol. 2015, pp. 1 – 12, 2015.

[21] A. C. Rencher and G. B. Schaalje, *Linear models in statistics*. John Wiley & Sons, 2008.

[22] W. Feller, *An introduction to probability theory and its applications. Vol. I*, 3rd ed. John Wiley & Sons Inc., 1968.

[23] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493 – 507, 1952.

[24] G. Barrera Vargas and P. Manrique-Mirón, "The asymptotic distribution of the condition number for random circulant matrices," *Extremes*, vol. 25, 2022.

[25] G. Casella and R. Berger, *Statistical Inference*. Cengage Learning, 2021.

[26] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*, ser. Springer Texts in Statistics. Springer, 2006.

[27] M. Leadbetter, G. Lindgren, and H. Rootzén, *Extremes and Related Properties of Random Sequences and Processes*. Springer, 1983.

[28] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Academic Press, Inc., 2008.

[29] L. Birgé and P. Massart, "Minimum contrast estimators on sieves: exponential bounds and rates of convergence," *Bernoulli*, vol. 4, no. 3, pp. 329–375, 1998.

[30] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, vol. 28, 2000.

V. Appendix

As a complement to what we derived for convnets in Proposition II.1, we show that the variance of the energy of a fully–connected layer with Gaussian initialization does not depend on the characteristics of the input signal. To see this, we use that any Gaussian matrix $\mathbf{W} \in \mathbb{R}^{M\times N}$ with $M \geq N$ is associated to a random tight frame of any order p, i.e., there is $C_p > 0$ such that $\mathbb{E}\left[\|\mathbf{W}\boldsymbol{x}\|^{2p}\right] = C_p\|\boldsymbol{x}\|^{2p}$ for any $p > 1$ [20]. For mean zero and variance σ^2 we have that $C_p = M(M+2)\cdots(M+2p-2)\sigma^{2p}$, see Example 4.4 in [20].

Proposition V.1. *Let $\boldsymbol{x} \in \mathbb{R}^N$ and $\mathbf{W} \in \mathbb{R}^{M\times N}$, $M \geq N$ be a random matrix with entries sampled i.i.d. from $\mathcal{N}(0,\sigma^2)$. Then*

$$\mathbb{E}\left[\|\mathbf{W}\boldsymbol{x}\|^2\right] = M\sigma^2\|\boldsymbol{x}\|^2, \tag{21}$$
$$\mathbb{V}\left[\|\mathbf{W}\boldsymbol{x}\|^2\right] = 2M\sigma^4\|\boldsymbol{x}\|^4. \tag{22}$$

Proof. For $p = 1$, we have that $C_1 = M\sigma^2$, showing (21). For the variance, we use that $C_2 = M(M+2)\sigma^4$ and deduce

$$\begin{aligned}\mathbb{V}\left[\|\mathbf{W}\boldsymbol{x}\|^2\right] &= \mathbb{E}\left[\left(\|\mathbf{W}\boldsymbol{x}\|^2 - M\sigma^2\|\boldsymbol{x}\|^2\right)^2\right] \\ &= \mathbb{E}\left[\|\mathbf{W}\boldsymbol{x}\|^4\right] - M^2\sigma^4\|\boldsymbol{x}\|^4 = 2M\sigma^4\|\boldsymbol{x}\|^4.\end{aligned}$$

■

By Proposition II.1, a random filterbank $\boldsymbol{\Phi}$ is a random tight frame of order one. For $p > 1$, this is in general not the case.

Proof of Lemma II.2. Given $\boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{w} \in \mathbb{R}^T$, we write the circulant convolution $\boldsymbol{x} * \boldsymbol{w}$ in Equation (1) as the matrix-vector multiplication $\mathbf{C}_T(x)\boldsymbol{w}$ where

$$\mathbf{C}_T(x) = \begin{pmatrix} \boldsymbol{x}[0] & \boldsymbol{x}[N-1] & \cdots & \boldsymbol{x}[N-T+1] \\ \boldsymbol{x}[1] & \boldsymbol{x}[0] & \cdots & \boldsymbol{x}[N-T+2] \\ \vdots & \vdots & & \vdots \\ \boldsymbol{x}[N-2] & \boldsymbol{x}[N-3] & \cdots & \boldsymbol{x}[N-T-1] \\ \boldsymbol{x}[N-1] & \boldsymbol{x}[N-2] & \cdots & \boldsymbol{x}[N-T] \end{pmatrix}$$

contains the first T columns of the circulant matrix generated by a reversed version of $\boldsymbol{x}$. The entries are given by

$$\mathbf{C}_T(x)[n,t] = \boldsymbol{x}[(n-t) \bmod N]$$

for $0 \leq n < N$ and $0 \leq t < T$. We write down its squared Euclidean norm as a quadratic form

$$\|\boldsymbol{x} * \boldsymbol{w}\|^2 = \langle \mathbf{C}_T(\boldsymbol{x})\boldsymbol{w}, \mathbf{C}_T(\boldsymbol{x})\boldsymbol{w}\rangle = \langle \boldsymbol{w}, \mathbf{Q}_T(\boldsymbol{x})\boldsymbol{w}\rangle$$

where $\mathbf{Q}_T(\boldsymbol{x}) = \mathbf{C}_T(\boldsymbol{x})^\top \mathbf{C}_T(\boldsymbol{x})$. Recalling the definition of circular autocorrelation (Equation 2), we conclude with

$$\begin{aligned}\mathbf{Q}_T(\boldsymbol{x})[t,t'] &= \sum_{n=0}^{N-1} \boldsymbol{x}[(n-t) \bmod N]\ \boldsymbol{x}[(n-t') \bmod N] \\ &= \mathbf{R}_{\boldsymbol{xx}}((t'-t) \bmod N).\end{aligned}$$

■

Proof of Lemma II.3. We apply Lemma II.2 with $0 \leq t < T$,

$$\mathbf{Q}_T(\boldsymbol{x})[t,t] = \mathbf{R}_{\boldsymbol{xx}}(0) = \sum_{n=0}^{N-1} \boldsymbol{x}[n]^2 = \|\boldsymbol{x}\|^2. \tag{23}$$

■

Proof of Proposition II.4. We recall Cantelli's inequality [22]:

$$\mathbb{P}\Big[Z - \mathbb{E}[Z] \geq \beta\Big] \leq \frac{\mathbb{V}[Z]}{\mathbb{V}[Z] + \beta^2}. \tag{24}$$

where $\beta > 0$ and Z has finite mean and variance. Given α and $\boldsymbol{x}$, we set $Z = \|\boldsymbol{\Phi}\boldsymbol{x}\|^2$ and $\beta = \alpha\|\boldsymbol{x}\|^2$. With Proposition II.1, we replace $\mathbb{E}[Z]$ by $JT\sigma^2\|\boldsymbol{x}\|^2$. With Lemma II.3, we replace $\|\boldsymbol{x}\|^4$ by $\mathbf{R}_{\boldsymbol{xx}}[0]^2$. Setting $\sigma^2 = (JT)^{-1}$ concludes the proof. ■

Our proof of Proposition II.5 hinges on the following lemma.

Lemma V.2 (Lemma 8 in Birgé *et al.* [29])**.** *For any $v, c, \beta > 0$,*

$$\inf_{\mu>0} \frac{\mu^2 v^2}{1-2\mu c} - \mu\beta \leq -\frac{\beta^2}{2c\beta + 2v^2}.$$

Proof of Proposition II.5. We show (10) via the generic Chernoff bounds for any random variable Z

$$\mathbb{P}[Z \geq \beta] \leq \inf_{\mu>0} \mathbb{E}\left[e^{\mu Z}\right] e^{-\mu\beta}. \tag{25}$$

We set $Z = \|\boldsymbol{\Phi}\boldsymbol{x}\|^2 - \|\boldsymbol{x}\|^2$ and use (13), together with Lemma II.3 to see that $Z = \sum_{k=0}^{T-1}\sum_{j=1}^{J} \sigma^2\lambda_k(\boldsymbol{y}_j[k]^2 - 1)$. A straightforward computation gives

$$\log \mathbb{E}\left[e^{\mu Z}\right] = \sum_{k=0}^{T-1}\sum_{j=1}^{J} \log \mathbb{E}\left[\exp\left(\mu\sigma^2\lambda_k(\boldsymbol{y}_j[k]^2 - 1)\right)\right].$$

Recall that $\frac{\boldsymbol{y}_j[k]}{\sigma^2} \sim \mathcal{N}(0,1)$. Analog to the proof of Lemma 1 in [30], we use that the mapping $\psi : u \mapsto \log \mathbb{E}\left[\exp\left(u\sigma^2(X^2-1)\right)\right]$ satisfies $\psi(u) \leq \frac{u^2\sigma^4}{1-2u\sigma^2}$ for any $X \sim \mathcal{N}(0,1)$ and $0 < u < \frac{1}{2\sigma^2}$. Since $\mathbf{C}_T(\boldsymbol{x})$ is a principal submatrix of a positive definite matrix (autocorrelation matrix), $\lambda_k > 0$ for all $k = 0, \ldots, T-1$. Therefore, for $\mu < \frac{1}{2\sigma^2 \max_k \lambda_k}$,

$$\log \mathbb{E}\left[e^{\mu Z}\right] \leq \sum_{k=0}^{T-1}\sum_{j=1}^{J} \frac{(\mu\lambda_k)^2\sigma^4}{1-2\mu\sigma^2\lambda_k} \leq \frac{\mu^2\sigma^4 J\|\boldsymbol{\lambda}\|_2^2}{1-2\mu\sigma^2\|\boldsymbol{\lambda}\|_\infty}. \tag{26}$$

Finally, using (26) and Lemma V.2 with $v^2 = \sigma^4 J\|\boldsymbol{\lambda}\|_2^2$ and $c = \sigma^2\|\boldsymbol{\lambda}\|_\infty$, we obtain

$$\begin{aligned}\inf_{\mu>0} \mathbb{E}\left[e^{\mu Z}\right] e^{-\mu\beta} &= \exp\left(\inf_{\mu>0} \log \mathbb{E}\left[e^{\mu Z}\right] - \mu\beta\right) \\ &\leq \exp\left(\inf_{\mu>0} \frac{\mu^2\sigma^4 J\|\boldsymbol{\lambda}\|_2^2}{1-2\mu\sigma^2\|\boldsymbol{\lambda}\|_\infty} - \mu\beta\right) \\ &\leq \exp\left(-\frac{\beta^2}{2\beta\sigma^2\|\boldsymbol{\lambda}\|_\infty + 2\sigma^4 J\|\boldsymbol{\lambda}\|_2^2}\right).\end{aligned}$$

Setting $\beta = \alpha\|\boldsymbol{x}\|^2$ and $\sigma^2 = (JT)^{-1}$ yields the claim. ■

Proof of Proposition III.2. Observe that

$$\min_{\|\boldsymbol{x}\|^2=1} \mathbf{R}_{\boldsymbol{xx}}(t)^2 = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise.} \end{cases} \quad \text{and} \quad \max_{\|\boldsymbol{x}\|^2=1} \mathbf{R}_{\boldsymbol{xx}}(t)^2 = 1$$

for $0 \leq t < T$. These extreme values are attained for an impulse and a constant signal respectively. Using these signals in Equation (4) of Proposition II.1 and etting $\sigma^2 = (TJ)^{-1}$ yields the result. ■