Maximum Entropy Models from Phase Harmonic Covariances

Sixin Zhang[1,4], Stéphane Mallat[1,2,3*]
[1] *ENS, PSL University, Paris, France*
[2] *Collège de France, Paris, France*
[3] *Flatiron Institute, New York, USA*
[4] *Center for Data Science, Peking University, Beijing, China*

November 25, 2019

Abstract

We define maximum entropy models of non-Gaussian stationary random vectors from covariances of non-linear representations. These representations are calculated by multiplying the phase of Fourier or wavelet coefficients with harmonic integers, which amounts to compute a windowed Fourier transform along their phase. Rectifiers in neural networks compute such phase windowing. The covariance of these harmonic coefficients capture dependencies of Fourier and wavelet coefficients across frequencies, by canceling their random phase. We introduce maximum entropy models conditioned by such covariances over a graph of local interactions. These models are approximated by transporting an initial maximum entropy measure with a gradient descent. The precision of wavelet phase harmonic models is numerically evaluated over turbulent flows and other non-Gaussian stationary processes.

Keywords— covariance, stationary process, phase, Fourier, wavelets, turbulence

1 Introduction

Many phenomena in physics, finance, signal processing and image analysis can be modeled as realizations of a non-Gaussian stationary process X. We often observe a single realization of dimension d from which the model must be estimated. Models of stochastic processes may be defined as maximum entropy distributions conditioned by a predefined family of moments [1, 2]. To estimate accurately these moments from a single or few realizations, the number of moments should not be too large. The main difficulty is to specify these moments so that the resulting maximum entropy model approximates well the distribution of the original process. We will show that the phase plays a crucial role to define moments that capture non-Gaussian properties.

[*] This work is supported by the ERC InvariantClass 320959

We introduce maximum entropy models based on covariances of a representation $\mathcal{R}(X)$ of X. Building a model amounts to choosing $\mathcal{R}$ and selecting a small subset of covariance coefficients, which are sufficient to compute an accurate maximum entropy model. This is interpreted as a graph model where covariance coefficients are kept along the edges of the graph. To estimate covariances from a single realization of X we use prior knowledge on symmetries of its probability distribution. If X is known to be stationary, then the group of symmetries includes translations but it may be larger, or may not include all translations if X is not stationary. The covariance is estimated from a single realization with an empirical average along the orbit of the group action. The number of covariance that must be estimated is typically reduced when the group of symmetries increase. This number must be much smaller than the dimension d of X to avoid introducing large estimation errors.

The central difficulty is to optimize the representation in order to build an accurate maximum entropy model. We shall see that the model error, measured with a Kullback-Leibler divergence, is equal to the excess of entropy of the maximum entropy model. To reduce this error, we define $\mathcal{R}$ so that $\mathcal{R}(X)$ is sparse, which leads to lower entropy models. We shall also impose that $\mathcal{R}$ is bi-Lipschitz continuous so that variations of $\mathcal{R}(X)$ are of the same order of magnitude as variations of X, which limits the variance of covariance estimations. Section 2 introduces bi-Lipschitz non-linear representations $\mathcal{R}$ computed from harmonics of phases of Fourier coefficients, to model non-Gaussian stationary processes. We then concentrate on phase harmonics of wavelet coefficients, which provide sparse representations and hence more accurate models of large classes of stationary processes.

To understand the importance of the phase, observe that if $\mathcal{R}(X)$ is the Fourier transform of a stationary X then its covariance matrix is diagonal. Fourier coefficients are uncorrelated at different frequencies because of stationary phase fluctuations [3]. If X is not Gaussian then Fourier coefficients are typically not independent. We modify $\mathcal{R}$ in order to capture this dependence. Section 2 explains that high order moments capture non-Gaussian statistics across frequencies by canceling random phase fluctuations. However, such high order moments have a large variance because z^k amplifies the variability of z if $|z|$ is large and $k > 1$. To preserve phase cancellation properties while avoiding large variance estimations, we replace $z^k = |z|^k \, e^{ik\varphi(z)}$ by phase harmonics $|z| \, e^{ik\varphi(z)}$ introduced in [4]. Phase harmonics keep the modulus and are therefore Lipschitz continuous. We show that they are obtained with a windowed Fourier transform along the phase. Section 2 studies the covariance of a Fourier phase harmonic operator $\mathcal{R}$.

Wavelet transforms of signals including singularities are often more sparse than Fourier transforms, which reduce errors of maximum entropy models. High order moments of wavelet coefficients have been used to characterize non-Gaussian multifractal properties of random processes and turbulent flows [5, 6]. As in the Fourier case, we replace these high order moments by phase harmonics, to define a Lipschitz continuous representation $\mathcal{R}$. Section 3 studies the covariance of this wavelet phase harmonic operator.

Wavelet phase harmonics are related to the pioneer work of Grossmann and Morlet [7], who showed qualitatively that complex wavelet transforms have a phase whose variations across scales provide important information on the geometry of transient structures. In image processing, Portilla and Simoncelli [8] brought a new perspective by showing that one can synthesize non-Gaussian image textures from the covariance of the modulus of wavelet coefficients and their phase at different scales. In an apparently different context, remarkable image texture synthesis were obtained by [9], by computing the covariance of output coefficients of a one-layer convolutional neural network, calculated with a rectifier. We show that these approaches correspond to different instantiations of phase harmonic covariances.

Section 4 studies maximum entropy models conditioned by empirical estimators of covariances. Sampling a maximum entropy distributions requires to use expensive algorithms which iterate over Gibbs samplers [10], which is not feasible over large size images. We thus rather use microcanonical models studied in [11], which transports an initial maximum entropy measure with a gradient descent which adjusts its moments. Section 5 defines low-dimensional foveal models of wavelet phase harmonics covariances. We evaluate the numerical precision of these models, to approximate non-Gaussian processes including turbulent flows. All calculations can be reproduced by a Python software available at `https://github.com/kymatio/phaseharmonics`.

Notations: We write z^* the complex conjugate of $z \in \mathbb{C}$, Y^* is the complex transpose of a matrix Y and $\mathcal{A}^*$ is the adjoint of an operator $\mathcal{A}$. The covariance of two random variables A and B is written $\mathrm{Cov}(A, B) = \mathbb{E}(A\, B^*) - \mathbb{E}(A)\, E(B)^*$. An inner product is written $\langle x, y\rangle = \int x(u)\, y(u)\, du$ in $\mathbf{L^2}(\mathbb{R}^d)$ and $\langle x, y\rangle = \sum_u x(u)\, y(u)$ in $\mathbb{R}^d$. The cardinal of a set S is $|S|$.

2 Fourier Phase Harmonic Representation

Next section introduces the general properties of maximum entropy models conditioned by covariance coefficients of a non-linear representation $\mathcal{R}$ over a graph. Section 2.2 then shows that the Fourier coefficients of a stationary process are uncorrelated because of phase fluctuations. It motivates the use of higher order moments to define a representation which can cancel random phase fluctuations and capture dependence across frequencies. Section 2.3 reviews the properties of phase harmonics, which also cancel the phase but defines a bi-Lipschitz representation as opposed to high order moments. Section 2.4 studies the resulting Fourier phase harmonic operator $\mathcal{R}$, whose covariance matrix captures dependencies of non-Gaussian random vectors across frequencies.

2.1 Maximum Entropy Covariance Graph Models

We introduce maximum entropy models of a stationary random vector X, conditioned by covariance coefficients of a representation $\mathcal{R}(X)$ which is also a random vector:

$$K_{\mathcal{R}} = \mathrm{Cov}(\mathcal{R}(X), \mathcal{R}(X)) = \mathbb{E}\Big((\mathcal{R}(X) - M_{\mathcal{R}})(\mathcal{R}(X) - M_{\mathcal{R}})^*\Big)$$

with $M_{\mathcal{R}} = \mathbb{E}(\mathcal{R}(X))$. We suppose that $X(u)$ is defined for u in a cube $\Lambda_d \subset \mathbb{Z}^r$ with d grid points, for example in $[1, d^{1/r}]^r$. If $r = 2$ then each realization is an image of d pixels. This section reviews general properties of maximum entropy models defined on a graph, which take advantage of known symmetries of the distribution of X.

Maximum entropy on a graph with symmetries Let us write $\mathcal{R}(X) = \{\mathcal{R}_v(X)\}_{v \in V}$ where V is a finite set of vertices in a graph that we now define. We introduce a maximum entropy model conditioned by the covariance between vertices

$$K_{\mathcal{R}}(v, v') = \mathrm{Cov}(\mathcal{R}_v(X), \mathcal{R}_{v'}(X)),$$

where v' is in a neighborhood $\mathcal{N}_v \subset V$ of v. We suppose that $v \in \mathcal{N}_v$. If $v' \in \mathcal{N}_v$ then $v \in \mathcal{N}_{v'}$ so it defines a reflexive and symmetric undirected graph (V, E), where the set of edges relates all neighbors $E = \{(v, v') : v \in V \, , \, v' \in \mathcal{N}_v\}$. The edge weights are the covariances $K_{\mathcal{R}}(v, v')$ over E. In statistical physics, $(\mathcal{R}_v(x) - M_{\mathcal{R}}(v))(\mathcal{R}_{v'}(x) - M_{\mathcal{R}}(v'))^*$ is called the interaction potential of $(v, v') \in E$, where the mean $M_{\mathcal{R}}$ is considered here as a predefined constant vector.

Let p be the probability density of X. We can reduce the constraints of the maximum entropy model if we also know a finite group G of symmetries of p. We consider linear unitary symmetries g from $\mathbb{R}^d$ to $\mathbb{R}^d$, which satisfy $p(g.x) = p(x)$ for all $x \in \mathbb{R}^d$. The stationarity of X in Λ_d means that p is invariant to periodic translations, that we write $g.x(u) = x(u - g)$, so G includes all translations. If p is invariant to rotations then G also includes rotations. Since p is invariant to the action of G, its covariance is also invariant to the action of any $g \in G$

$$\mathrm{Cov}(\mathcal{R}_v(g.X), \mathcal{R}_{v'}(g.X)) = \mathrm{Cov}(\mathcal{R}_v(X), \mathcal{R}_{v'}(X)).$$

Let $|G|$ be the total number of symmetries. These $|G|$ conditions will be included in the maximum entropy model.

The entropy of a density $\tilde{p}$ on $\mathbb{R}^d$ is

$$H(\tilde{p}) = -\int \tilde{p}(x) \, \log \tilde{p}(x) \, dx.$$

A maximum entropy macrocanonical model $\widetilde{X}$ conditioned by the covariance $K_{\mathcal{R}}$ over E and by the symmetry group G has a probability density $\tilde{p}$ which maximizes $H(\tilde{p})$ and satisfies covariance moment conditions for all $(v, v', g) \in E \times G$

$$\int (\mathcal{R}_v(g.x) - M_{\mathcal{R}}(v)) \, (\mathcal{R}_{v'}(g.x) - M_{\mathcal{R}}(v'))^* \, \tilde{p}(x) \, dx = K_{\mathcal{R}}(v, v'). \tag{1}$$

If there exists a solution to this convex optimization with equality constraints then it is a unique and it can then be written [2]

$$\tilde{p}(x) = Z^{-1} \, \exp\Big(-\frac{1}{|G|} \sum_{(v,v',g) \in E \times G} \beta_{v,v'} \, (\mathcal{R}_v(g.x) - M_{\mathcal{R}}(v)) \, (\mathcal{R}_{v'}(g.x) - M_{\mathcal{R}}(v'))^*\Big), \tag{2}$$

where $\beta_{v,v'}$ is the Lagrange multiplier associated to each equality condition (1) and does not depend upon g. The sum in the exponential is the Gibbs energy, and Z is the partition function. Since $K_\mathcal{R}(v,v') = K^*_\mathcal{R}(v',v)$ we have $\beta_{v,v'} = \beta^*_{v',v}$. We verify that for all $g \in G$, $\tilde{p}(g.x) = \tilde{p}(x)$ so G is also a group of symmetries of $\tilde{p}$. If G includes all translations then $\tilde{p}$ is stationary.

In the particular case where all $\mathcal{R}_v(x)$ are linear operators then the Gibbs energy is bilinear and $\tilde{p}(x)$ is thus a Gaussian distribution. Appendix A shows that the Lagrange multipliers can then be computed efficiently [2]. If some of the $\mathcal{R}_v(x)$ are non-linear then computing the Lagrange multipliers is computationally very expensive when the dimension d of X and the total number $|E|$ of moments is large.

Covariance estimation from symmetries The covariance estimation from a single realization $\bar{x}$ of X is calculated with an average over the symmetries of G. The orbit of the action of G on $\bar{x}$ is the set of $\{g.\bar{x}\}_{g\in G}$. The empirical estimation of $M_\mathcal{R}(v) = \mathbb{E}(R_v(X))$ is computed as an empirical average over this orbit

$$\widetilde{M}_\mathcal{R} = \frac{1}{|G|}\sum_{g\in G}\mathcal{R}(g.\bar{x}). \tag{3}$$

Similarly, $K_\mathcal{R}$ is estimated with an average on the same orbit

$$\widetilde{K}_{\mathcal{R}\bar{x}} = \frac{1}{|G|}\sum_{g\in G}\Big(\mathcal{R}(g.\bar{x}) - \widetilde{M}_\mathcal{R}\Big)\Big(\mathcal{R}(g.\bar{x}) - \widetilde{M}_\mathcal{R}\Big)^*. \tag{4}$$

Summing over all transformations of $\bar{x}$ by $g \in G$ is called "data augmentation" in machine learning. The larger $|G|$ the more accurate the estimation. The maximum entropy model is estimated from a single realization $\bar{x}$ by replacing the mean $M_\mathcal{R}$ and covariance $K_\mathcal{R}$ in (1) by their estimation $\widetilde{M}_\mathcal{R}$ and $\widetilde{K}_{\mathcal{R}\bar{x}}$.

The existence of symmetries also reduces the number of covariances that must to be estimated. We define E_G as a minimum set of edges (v,v') such that for any $(v,v') \in E$ there exists $(v_1, v_1', g) \in E_G \times G$ such that $\mathcal{R}_v(x) = \mathcal{R}_{v_1}(g.x)$ and $\mathcal{R}_{v'}(x) = \mathcal{R}_{v_1'}(g.x)$. Since p is invariant to the action of G, its covariance $K_\mathcal{R}$ is also invariant:

$$K_\mathcal{R}(v,v') = \mathrm{Cov}(\mathcal{R}_{v_1}(g.X), \mathcal{R}_{v_1'}(g.X)) = K_\mathcal{R}(v_1, v_1').$$

It is therefore sufficient to estimate covariance coefficients indexed by E_G to specify all covariances indexed by E. The set E_G is interpreted as a set of sufficient statistics. Given a realization $\bar{x}$ of dimension d, to control the overall covariance estimation errors we must insure that $|E_G| \ll d$.

Bi-Lipschitz continuity The estimation of covariance coefficients may have a large variance in the presence of rare outliers. These outliers induce a large variability in the empirical sum (4) depending upon the realization $\bar{x}$ of X. To avoid amplifying these outliers, we impose that $\mathcal{R}$ is bi-Lipschitz, so that the variations of $\mathcal{R}(X)$ are of the same order as the variations of X. It also insures that $\mathcal{R}$ is an invertible operator.

The representation $\mathcal{R}$ is bi-Lipschitz if there exists $A_\mathcal{R} > 0$ and $B_\mathcal{R}$ such that for all $(x,x') \in \mathbb{R}^{2d}$

$$A_\mathcal{R}\,\|x - x'\|^2 \le \|\mathcal{R}(x) - \mathcal{R}(x')\|^2 \le B_\mathcal{R}\,\|x - x'\|^2. \tag{5}$$

For any random vector Z, we write $\sigma^2(Z) = \mathbb{E}(\|Z - \mathbb{E}(Z)\|^2)$, which is the trace of its covariance. The following proposition proves that the variance of $\mathcal{R}(X)$ and X have the same order of magnitude.

Proposition 2.1. *If $\mathcal{R}$ is bi-Lipschitz then*

$$A_{\mathcal{R}}\, \sigma^2(X) \leq \sigma^2(\mathcal{R}(X)) \leq B_{\mathcal{R}}\, \sigma^2(X). \tag{6}$$

Proof: If X' and X are two independent random vectors having same probability distribution then the bi-Lipschitz bounds (5) of $\mathcal{R}$ imply that

$$A_{\mathcal{R}}\, \mathbb{E}(\|X - X'\|^2) \leq \mathbb{E}(\|\mathcal{R}(X) - \mathcal{R}(X')\|^2) \leq B_{\mathcal{H}}\, \mathbb{E}(\|X - X'\|^2). \tag{7}$$

If Y and Y' are two i.i.d random variables then we verify that

$$\mathbb{E}(|Y - Y'|^2) = 2\, \mathbb{E}(|Y - \mathbb{E}(Y)|^2).$$

where the first expected value is taken relatively to the joint distribution of Y and Y'. It results that $\mathbb{E}(\|X - \mathbb{E}(X)\|^2) = 2^{-1}\, \mathbb{E}(\|X - X'\|^2)$ and $\mathbb{E}(\|\mathcal{R}(X) - \mathbb{E}(\mathcal{R}(X))\|^2) = 2^{-1}\, \mathbb{E}(\|\mathcal{R}(X) - \mathcal{R}(X'))\|^2)$. Inserting these equalities in (7) proves proves (6). □

Maximum entropy reduction with sparsity Zhu, Wu and Mumford [12] have proposed to optimize maximum entropy parameterized models by minimizing the resulting maximum entropy. Indeed, if X has a probability density p then model error can be evaluated with a Kullback-Leibler divergence

$$D_{KL}(p||\tilde{p}) = \int p(x)\, \log \frac{p(x)}{\tilde{p}(x)}\, dx.$$

We verify that $\int p(x)\, \log \tilde{p}(x)\, dx = \int \tilde{p}(x)\, \log \tilde{p}(x)\, dx$ by inserting (2) and by using the equalities (1). Since $H(p) = \int p(x)\, \log p(x)\, dx$, it results that the Kullback-Leibler divergence is equal to the excess of entropy of the maximum entropy model:

$$D_{KL}(p||\tilde{p}) = H(\tilde{p}) - H(p) \geq 0. \tag{8}$$

We thus reduce the model error by reducing the maximum entropy $H(\tilde{p})$. The minimum $H(\tilde{p}) = H(p)$ is reached if and only if $\tilde{p} = p$.

In our case, the model depends upon the choice of $\mathcal{R}$ and of the edges $E \subset V^2$ of the covariance graph. Optimizing $\mathcal{R}$ by calculating $H(\tilde{p})$ is computationally too expensive. However, we explain below that we can minimize an upper bound of $H(\tilde{p})$ by finding a representation such that $\mathcal{R}(X)$ is as sparse as possible, which gives a partial control on the maximum entropy. Let $\widetilde{X}$ be a maximum entropy model of density $\tilde{p}$. If $\mathcal{R}$ has a stable inverse, which we guarantee with the bi-Lipschitz condition, then an upper bound of $H(\tilde{p})$ can be computed from the entropy of each marginal $\mathcal{R}_v(\widetilde{X})$. Such an entropy is small if $\mathcal{R}_v(\widetilde{X})$ is sparse, because it then has a narrow probability density centered at 0. To impose this sparsity we must find a representation $\mathcal{R}$ such that $\mathcal{R}_v(X)$ is sparse. This sparsity must also be captured by diagonal covariance coefficients $K_{\mathcal{R}}(v,v)$ so that the maximum entropy model $\widetilde{X}$, conditioned by these moments retains this sparsity. This is the case for Fourier and wavelet phase harmonic representations studied in Sections 2.4 and 3.2. For non-Gaussian processes, it amounts to represent "coherent structures" with as few non-zero coefficients as possible.

Increasing the number $|E|$ of moment conditions can further reduce $H(\tilde{p})$ but it also increases the statistical error when estimating these moments from a single realization of X. The choice of E is thus a trade-off between the model error (bias) and the estimation error (variance).

2.2 Fourier Phase and High Order Moments

The covariance of a stationary random vector is diagonalized by the discrete Fourier transform, because of random phase fluctuations. This suggests defining a covariance representation in a Fourier basis. For non-Gaussian processes, the dependence of Fourier coefficients is partly captured by high order moments which cancel random phase fluctuations.

For $\mathcal{R} = Id$, since $X(u)$ for $u \in \Lambda_d$ is stationary, $K_{\mathcal{R}}(u,u') = \mathrm{Cov}(X(u), X(u'))$ only depends on $u - u'$. We write $|u|$ the norm of $u \in \Lambda_d$ and $u'.u$ the inner product. A low-dimensional maximum entropy model is constructed on $V = \Lambda_d$ by restricting covariances to neighborhoods of fixed radius: $\mathcal{N}_v = \{v' : |v - v'| \leq c\}$. The radius c is chosen to be large enough to capture long range correlations. Since X is stationary, the symmetry group includes translations and we can thus define E_G by setting $v = 0$. The maximum entropy model then defines a Gauss-Markov random vector [2] specified by covariances in a small neighborhood. This model does not capture non-Gaussian properties. If Λ_d is an r-dimensional grid, then the model size is $|E_G| = O(c^r)$. It may be large if the process has long-range spatial correlations.

To better understand how to capture non-Gaussian properties, we study these covariance coefficients in a Fourier basis. We write $\widehat{x} = \mathcal{F}_u x$ the discrete Fourier transform of x over Λ_d:

$$\widehat{x}(\omega) = \sum_{u \in \Lambda_d} x(u)\, e^{-i\omega.u} \quad \text{for} \quad \omega = 2\pi d^{-1/r} m \quad \text{with} \quad m \in \Lambda_d. \tag{9}$$

The Fourier representation $\mathcal{R} = \mathcal{F}_u$ is indexed by $\omega \in V = 2\pi d^{-1/r}\Lambda_d$. The covariance for $(\omega, \omega') \in V^2$ is

$$K_{\mathcal{R}}(\omega, \omega') = \mathrm{Cov}(\widehat{X}(\omega)\, \widehat{X}(\omega')).$$

If $\omega \neq 0$ then $\mathbb{E}(\widehat{X}(\omega)) = 0$. If $\omega \neq \omega'$ then $K_{\mathcal{R}}(\omega, \omega') = 0$ because of random phase fluctuations. Indeed, translating $X(u)$ by any $\tau \in G_q$ multiplies $\widehat{X}(\omega)$ by $e^{-i\tau.\omega}$. Since X is stationary, this translation is a symmetry which does not modify $\mathrm{Cov}(\widehat{X}(\omega), \widehat{X}(\omega'))$. It results that

$$\mathrm{Cov}(\widehat{X}(\omega), \widehat{X}(\omega')) = e^{i(\omega-\omega').\tau}\, \mathrm{Cov}(\widehat{X}(\omega), \widehat{X}(\omega')). \tag{10}$$

Since this is true for any $\tau \in \Lambda_d$ it implies that

$$\mathrm{Cov}(\widehat{X}(\omega)\, \widehat{X}(\omega')) = 0 \quad \text{if} \quad \omega \neq \omega'. \tag{11}$$

Since the discrete Fourier transform is periodic beyond $[0, 2\pi]^r$, any equality or non-equality between frequencies must be understood modulo 2π along the r directions. If X is Gaussian then $\widehat{X}$ is also Gaussian so this non-correlation implies that $\widehat{X}(\omega)$ and $\widehat{X}(\omega')$ are independent. However, if X is non-Gaussian then $\widehat{X}(\omega)$ and $\widehat{X}(\omega')$ are typically not independent.

Phase cancellation with higher order moments To capture the dependencies of Fourier coefficients across frequencies, one can use high order moments [13]. For any $k \in \mathbb{Z}$, similarly to (10), translating X by $\tau \in \Lambda_d$ yields

$$\mathrm{Cov}(\widehat{X}(\omega)^k, \widehat{X}(\omega')^{k'}) = e^{i(k\omega - k'\omega').\tau} \, \mathrm{Cov}(\widehat{X}(\omega)^k, \widehat{X}(\omega')^{k'}). \tag{12}$$

A high-order Fourier representation $\mathcal{R}_v(X) = \widehat{X}(\omega)^k$ is indexed by $v = (\omega, k) \in V$ with $0 \leq k \leq k_{\max}$. It results from (12) that

$$K_{\mathcal{R}}(v, v') = \mathrm{Cov}(\widehat{X}(\omega)^k, \widehat{X}(\omega')^{k'}) = 0 \quad \text{if} \quad k\omega \neq k'\omega' \ . \tag{13}$$

If $k\omega = k'\omega'$ and X is not Gaussian then this covariance is typically non-zero because the phase variations of $\widehat{X}(\omega')^{k'*}$ cancel the phase variations of $\widehat{X}(\omega)^k$. This is also the key idea behind the use of bi-spectrum moments [14]. By adjusting (k, k') these moments provide some dependency information between $\widehat{X}(\omega)$ and $\widehat{X}(\omega')$ for $\omega \neq \omega'$.

For example, let $X(u) = x(u - S)$ be a random shift vector, where $x(u)$ is a fixed signal supported in Λ_d and S is a random periodic shift which is uniformly distributed in Λ_d. It is a stationary process whose Fourier coefficients have a random phase: $\widehat{X}(\omega) = \widehat{x}(\omega) \, e^{-iS\omega}$. In this case, if $k\omega = k'\omega'$ then

$$\mathrm{Cov}(\widehat{X}(\omega)^k, \widehat{X}(\omega')^{k'}) = \widehat{x}(\omega)^k \, \widehat{x}(\omega')^{*k}.$$

It is non-zero at frequencies where $\widehat{x}$ does not vanish. This shows that covariances of the high order Fourier exponents $\mathcal{R}(X)$ can capture the dependence of Fourier coefficients at different frequencies. However, this high order Fourier representation is not Lipschitz. Indeed, when $k > 1$, the exponent k amplifies the variability of each random variables $\widehat{X}(\omega)$, so estimators of covariance coefficients have a large variance.

2.3 Phase Windowed Fourier Transform and Harmonics

By replacing high order exponents by an exponent on the phase only, we show the resulting phase harmonic operator is bi-Lipschitz. These coefficients are obtained by applying a windowed Fourier transform along phases. We prove that rectifiers in neural networks compute windowed transformations on phases.

For $z = |z| e^{i\varphi(z)} \in \mathbb{C}$, $z^k = |z|^k \, e^{ik\varphi(z)}$ exponentiates the modulus and phase together. It is the modulus exponentiation which amplifies the variance of a random variable although it has little role in ensuring that correlation of Fourier coefficients are non-zero. We thus eliminate the modulus exponentiation and we replace it by the phase harmonics introduced in [4]. A phase harmonic computes a power $k \in \mathbb{Z}$ of the phase only:

$$[z]^k = |z| \, e^{ik\varphi(z)} \ .$$

It preserves the modulus: $|[z]^k| = |z|$. We shall see that it is computed with a windowed Fourier transform on the phase of z. This section reviews the properties of this phase windowed Fourier transform introduced in [4]. Next section defines a new representation $\mathcal{R}$ by applying it to Fourier coefficients.

Phase windowed Fourier transform A windowed Fourier transform of a signal $x(u)$ is a linear operator which multiplies $x(u)$ by a translated window along u and computes a Fourier transform of the windowed signal along u. Because the phase $\varphi(z)$ is a non-linear function of $z \in \mathbb{C}$, a windowed Fourier transform on the phase is a non-linear operator.

The phase of z is translated by a variable $\alpha \in [0, 2\pi]$, and its support is limited by a 2π periodic window $h(\alpha)$:

$$\mathcal{H}(z) = \{|z|\, h(\varphi(z) + \alpha)\}_{\alpha \in [0,2\pi]}. \tag{14}$$

This phase windowing is non-linear. A phase windowed Fourier transform computes the Fourier transform of $\mathcal{H}(z)$ relatively to α. Let us write $\widehat{h} = \mathcal{F}_\alpha(h)$ the Fourier transform along phases:

$$\widehat{h}(k) = \frac{1}{2\pi} \int_0^{2\pi} h(\alpha)\, e^{-ik\alpha}\, d\alpha. \tag{15}$$

Applying $\mathcal{F}_\alpha$ to (14) gives

$$\widehat{\mathcal{H}}(z) = \{\widehat{h}(k)\, [z]^k\}_{k \in \mathbb{Z}}\ . \tag{16}$$

It proves that a phase windowed Fourier transform $\widehat{\mathcal{H}} = \mathcal{F}_\alpha\, \mathcal{H}$ computes weighted phase harmonics. The harmonic weights $\widehat{h}(k)$ amplify or eliminate different phase harmonics. The more regular the phase window h the faster the decay of harmonic weights.

A rectifier $\rho(a) = \max(a, 0)$ is an important example of non-linearity which acts as a phase windowing. Indeed

$$\rho(\mathrm{Real}(z)) = |z|\, \rho(\cos\varphi(z))\ ,$$

so

$$\{\rho(\mathrm{Real}(e^{i\alpha} z))\}_{\alpha \in [0,2\pi]} = \mathcal{H}(z) \ \text{ with } \ h(\alpha) = \rho(\cos\alpha). \tag{17}$$

The rectifier phase window $\rho(\cos\alpha)$ is positive and supported in $[-\pi/2, \pi/2]$. The corresponding harmonic weights are computed in [4] with the Fourier integral (15):

$$\widehat{h}(k) = \begin{cases} \frac{-(i)^k}{\pi(k-1)(k+1)} & \text{if } k \text{ is even} \\ \frac{1}{4} & \text{if } k = \pm 1 \\ 0 & \text{if } |k| > 1 \text{ is odd} \end{cases}\ . \tag{18}$$

Their decay is slow because $h(\alpha)$ has discontinuous derivatives at $\pm\pi/2$.

Bi-Lipschitz continuity We mentioned that polynomial exponents amplify the variability of random variables around their mean. Indeed if $(z,z') \in \mathbb{C}^2$ then $|z^k - z'^k|/|z-z'|$ may be arbitrarily large if $k > 1$. On the contrary, a phase harmonic preserves the modulus which provides a bound on such amplification. It is proved in [4] that it is Lipschitz continuous

$$\forall (z,z') \in \mathbb{C}^2 \quad , \quad |[z]^k - [z']^k| \leq \max(|k|,1)\, |z-z'|. \tag{19}$$

The distance $|z-z'|$ is therefore amplified by at most $|k|$.

This Lipschitz continuity is extended to phase windowed Fourier transforms, which are shown to be bi-Lipschitz. The Fourier transform preserves norms and distances. Calculating the norm of (16) gives

$$\|\mathcal{H}(z)\| = \|\widehat{\mathcal{H}}(z)\| = \|h\|^2\, |z|^2,$$

with $\|h\|^2 = \sum_k |\widehat{h}(k)|^2$. Following [4], we also derive from (19) that a phase windowed Fourier transform is bi-Lipschitz over complex numbers $(z,z') \in \mathbb{C}^2$

$$A_{\mathcal{H}}\, |z-z'|^2 \leq \|\widehat{\mathcal{H}}(z) - \widehat{\mathcal{H}}(z)'\|^2 \leq B_{\mathcal{H}}\, |z-z'|^2 \ , \tag{20}$$

with $A_{\mathcal{H}} = 2\,|\widehat{h}(1)|^2$ and $B_{\mathcal{H}} = |\widehat{h}(0)|^2 + \sum_{k\in\mathbb{Z}} k^2\, |\widehat{h}(k)|^2$. The upper bound $B_{\mathcal{H}}$ is derived from (19) and the lower bound is obtained isolating the terms corresponding to $k = \pm 1$. For the rectifier filter (18), we get $A_{\mathcal{H}} = 1/8$ and $B_{\mathcal{H}} = 1/4 + 1/\pi^2$. The following theorem gives a tight upper Lipschitz constant for the rectifier filter, proved in Appendix B.

Theorem 2.1. *The rectifier phase window $h(\alpha) = \max(\cos\alpha, 0)$ has a Lipschitz upper-bound $B_{\mathcal{H}} = 1/4$.*

2.4 Phase Harmonics of Fourier Coefficients

Section 2.2 explains that if X is stationary then $\widehat{X}(\omega)$ and $\widehat{X}(\omega')$ are not correlated if $\omega \neq \omega'$. Similarly to high order moments, we show that $[\widehat{X}(\omega)]^k$ and $[\widehat{X}(\omega')]^{k'}$ may become correlated because of random phase cancellations, and it defines a sparse covariance matrix.

Applying $\widehat{\mathcal{H}}$ to each Fourier coefficients gives

$$\widehat{\mathcal{H}}(\widehat{X}(\omega)) = \left\{ \widehat{h}(k)\, [\widehat{X}(\omega)]^k \right\}_{k\in\mathbb{Z}} .$$

It defines a non-linear representation $\mathcal{R} = \widehat{\mathcal{H}}\mathcal{F}_u$ indexed by $v = (\omega, k)$. Since $\widehat{\mathcal{H}}$ is bi-Lipschitz and the Fourier transform is unitary up to a factor d, it results that $\mathcal{R} = \widehat{\mathcal{H}}\mathcal{F}_u$ is bi-Lipschitz with Lipschitz constants $A_{\mathcal{R}} = dA_{\mathcal{H}}$ and $B_{\mathcal{R}} = dB_{\mathcal{H}}$. Proposition 2.1 implies that

$$dA_{\mathcal{H}}\, \sigma^2(X) \leq \sigma^2(\widehat{\mathcal{H}}(X)) \leq dB_{\mathcal{H}}\, \sigma^2(X). \tag{21}$$

Sparse phase harmonic covariance Phase harmonic covariance coefficients are

$$K_{\widehat{\mathcal{H}}\mathcal{F}_u}(\omega, k, \omega', k') = \widehat{h}(k)\, \widehat{h}(k')^*\, \mathrm{Cov}([\widehat{X}(\omega)]^k\,,\, [\widehat{X}(\omega')]^{k'}).$$

If $\widehat{h}$ is compactly supported in $[0, k_{\max}]$, since X is of dimension d, the covariance $K_{\widehat{\mathcal{H}}\mathcal{F}}$ has $(k_{\max}+1)^2 d^2$ coefficients. The following proposition proves that coefficients of $K_{\widehat{\mathcal{H}}\mathcal{F}}$ are mostly zero and it is diagonal if X is Gaussian. Diagonal values are specified.

Theorem 2.2. *If X is real stationary then*

$$\mathrm{Cov}([\widehat{X}(\omega)]^k, [\widehat{X}(\omega')]^{k'}) = 0 \quad \textit{if} \;\; k\omega \neq k'\omega'. \tag{22}$$

Along the diagonal, if $\omega \neq 0$ and $k \neq 0$ or if $\omega = 0$ and k is odd then

$$\mathrm{Cov}([\widehat{X}(\omega)]^k, [\widehat{X}(\omega)]^k) = \mathrm{Cov}(\widehat{X}(\omega), \widehat{X}(\omega))\ . \tag{23}$$

If $\omega = 0$ and k is even then

$$\mathrm{Cov}([\widehat{X}(0)]^k, [\widehat{X}(0)]^k) = \mathrm{Cov}(|\widehat{X}(0)|, |\widehat{X}(0)|)\ . \tag{24}$$

For all $\omega \neq 0$

$$\frac{\mathrm{Cov}(|\widehat{X}(\omega)|, |\widehat{X}(\omega)|)}{\mathrm{Cov}(\widehat{X}(\omega), \widehat{X}(\omega))} = 1 - \frac{\mathbb{E}(|\widehat{X}(\omega)|)^2}{\mathbb{E}(|\widehat{X}(\omega)|^2)}. \tag{25}$$

If X is Gaussian then $K_{\widehat{\mathcal{H}}\mathcal{F}}$ is diagonal and $\mathbb{E}(|\widehat{X}(\omega)|)^2/\mathbb{E}(|\widehat{X}(\omega)|^2) = \pi/4$ if $\omega \neq 0$.

The proof is in Appendix C. All equalities or non-equalities on frequencies must be understood modulo 2π in dimension r. Property (22) proves that $K_{\widehat{\mathcal{H}}\mathcal{F}_u}$ is highly sparse. If X is not Gaussian then off-diagonal coefficients are typically not zero when $k\omega = k'\omega'$. For $k = k' = 0$, we get d^2 modulus covariances $\mathrm{Cov}(|\widehat{X}(\omega)|, |\widehat{X}(\omega')|)$ which are a priori non-zero for all (ω, ω'). It provides no information on the phase of $\widehat{X}(\omega)$ and $\widehat{X}(\omega')$. Phase correlations are captured when $k \neq 0$. If $\omega \neq 0$, it results from (22) that $\mathrm{Cov}([\widehat{X}(\omega)]^k, [\widehat{X}(\omega')]^{k'}) \neq 0$ only if ω' is colinear with ω. For a grid Λ_d of size d, there are $O(d^{1/r})$ such frequencies ω'. The total number of non-zero off-diagonal covariance coefficients is thus at most $d^2 + O(d^{1+1/r})$.

For a non-Gaussian process, off-diagonal coefficients with $k\omega = k'\omega'$ are typically non-zero. This is illustrated with a random shift process $X(u) = x(u - S)$ where S is uniformly distributed in Λ_d. If $k\omega = k'\omega'$ then one can verify that

$$\mathrm{Cov}(\widehat{X}(\omega)^k, \widehat{X}(\omega')^{k'}) = [\widehat{x}(\omega)]^k\, [\widehat{x}(\omega')]^{-k'} \;\text{ if }\; \omega \neq 0, \omega' \neq 0.$$

If $\hat{x}(\omega)$ does not vanish then all these coefficients are non-zero.

Along the diagonal, (23) and (24) prove that all coefficients are either equal to $\mathrm{Cov}(\widehat{X}(\omega), \widehat{X}(\omega))$ or to $\mathrm{Cov}(|\widehat{X}(\omega)|, |\widehat{X}(\omega)|)$. Property (25) also proves that the ratio between these covariance values depend upon the ratio $\mathbb{E}(|\widehat{X}(\omega)|)^2/\mathbb{E}(|\widehat{X}(\omega)|^2)$. Since $\mathbb{E}(X(\omega)) = 0$ for $\omega \neq 0$, this last ratio measures the sparsity of the random value $\widehat{X}(\omega)$. If it is smaller than the Gaussian ratio $\pi/4$ then $\widehat{X}(\omega)$ has a high probability to be relatively small and it has large amplitude outliers of low probability. The probability distribution of $\widehat{X}(\omega)$ is then highly peaked in zero with a long tail. The smaller this ratio the lower the entropy of this marginal probability distribution.

We saw in (8) that a maximum entropy model $\widetilde{X}$ introduces errors if its entropy is much larger than the entropy of X. The entropy $H(\tilde{p})$ is bounded by the sum of the entropy of each random variable $\widehat{X}(\omega)$, which is conditioned by $\mathrm{Cov}(\widehat{X}(\omega), \widehat{X}(\omega)$ and $\mathrm{Cov}(|\widehat{X}(\omega)|, |\widehat{X}(\omega)|)$. It gives an upper bound on the entropy $H(\tilde{p})$, which decreases if the sparsity of $\widehat{X}(\omega)$ increases. However, if $X(u)$ has sharp localized transitions then its Fourier coefficients have typically a large amplitude across most frequencies and are not sparse. On the contrary, wavelet coefficients may be sparse. In this case, we shall capture this sparsity and thus compute maximum entropy models of lower entropy, by replacing the Fourier transform by a wavelet transform.

3 Wavelet Phase Harmonics

To model random vectors whose realizations include singularities and sharp transitions, we replace the Fourier transform by a wavelet transform. Wavelet transform can provide sparse representations of such signals, and wavelet coefficients may have a short range dependence allowing to define low-dimensional models. We concentrate on two-dimensional image applications. Section 3.1 reviews the construction of complex steerable wavelet frames and the properties of their covariance matrices. Section 3.2 studies the covariance of wavelet phase harmonics.

3.1 Steerable Wavelet Frame Covariance

Complex steerable wavelet frames were introduced in [15] and are further studied in [16], to easily compute wavelet coefficients of rotated images. For simplicity, we begin by introducing wavelets as a localized functions $\psi(u)$ for $u \in \mathbb{R}^2$, with $\int \psi(u)\, du = 0$. Complex steerable wavelets have a Fourier $\widehat{\psi}(\omega)$ concentrated over one-half of the Fourier domain. We impose that $\psi(-u) = \psi^*(u)$ so that $\widehat{\psi}(\omega)$ is real. This Fourier transform is centered at a frequency $\xi \in \mathbb{R}^2$ and is non negligible for $\omega \in \mathbb{R}^2$ such that $|\omega - \xi| \leq C'|\xi|$ for some $C' > 0$. Figure 1 gives an example of such a wavelet, which is specified in Appendix D.

Let r_ℓ be a rotation by an angle $2\ell\pi/L$. Multiscale steerable wavelets are derived from ψ with dilations by 2^j for $j \in \mathbb{Z}$, and rotations over L angles $\theta = 2\ell\pi/L$ for $0 \leq \ell < L$

$$\psi_\lambda(u) = 2^{-j}\psi(2^{-j}r_{-\ell}u) \quad \Rightarrow \quad \widehat{\psi}_\lambda(\omega) = 2^j\widehat{\psi}(2^j r_\ell \omega) \;\text{ with }\; \lambda = 2^{-j}r_{-\ell}\,\xi. \tag{26}$$

Since $\widehat{\psi}(\omega)$ is non negligible for $|\omega - \xi| \leq C|\xi|$ it results that $\widehat{\psi}_\lambda(\omega)$ is centered at λ and non negligible for $|\omega - \lambda| \leq C|\lambda|$. In space, $\psi_\lambda(u)$ is non-negligible for $|u| \leq C'|\lambda|^{-1}$. We limit the scale 2^j to a maximum 2^J. The lowest frequencies are captured by a wavelet centered at $\lambda = 0$. It is computed by dilating a function $\phi(u)$ such that $\int \phi(u)\, du = 1$:

$$\psi_0(u) = 2^{-J}\phi(2^{-J}u) \quad \Rightarrow \quad \widehat{\psi}_0(\omega) = 2^J\widehat{\phi}(2^J\omega). \tag{27}$$

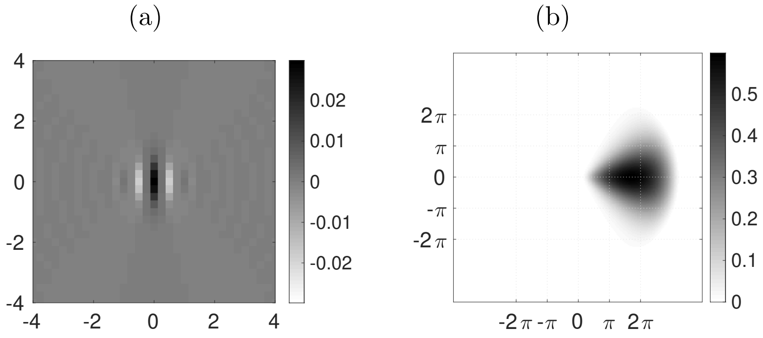

Figure 1: (a): Real-part of steerable bump wavelet $\psi(u)$. (b): Fourier transform $\widehat{\psi}(\omega)$.

A wavelet frame is constructed by translating each ψ_λ for $\lambda \neq 0$ by $u = 2^{j-1}n$ and ψ_0 by $u = 2^{J-1}n$ for all $n \in \mathbb{Z}^2$. It introduces a factor 2 oversampling relatively to a wavelet orthonormal basis [17], which creates some redundancy. The wavelet transform of $x \in \mathbf{L}^2(\mathbb{R}^2)$ is defined by

$$\mathcal{W}x = \{x \star \psi_\lambda(u)\}_{(\lambda,u)\in\Gamma}$$

where Γ is a frequency-space index set with $(\lambda, u) = (2^{-j} r_{-\ell}\xi, 2^{j-1}n)$ for $1 \leq j \leq J$, $0 \leq \ell < L$, $n \in \mathbb{Z}$, or $(\lambda, u) = (0, 2^{J-1}n)$.

Under appropriate conditions on ψ, the wavelet family $\{\psi_\lambda(\cdot - u)\}_{(\lambda,u)\in\Gamma}$ is a frame of $\mathbf{L}^2(\mathbb{R}^2)$ [16]. This means that there exists $0 < A_\mathcal{W} \leq B_\mathcal{W}$ such that for any $x \in \mathbf{L}^2(\mathbb{R}^2)$

$$A_\mathcal{W}\, \|x\|^2 \leq \|\mathcal{W}x\|^2 \leq B_\mathcal{W}\, \|x\|^2 \tag{28}$$

with $\|\mathcal{W}x\|^2 = \sum_{(\lambda,u)\in\Gamma} |x \star \psi_\lambda(u)|^2$.

The wavelet transform can be redefined over discrete images x of d pixels supported in a two-dimensional square grid Λ_d, uniformly sampled at intervals 1 in $[1, d^{1/2}]^2$. It requires to discretize and modify "boundary wavelets" whose supports intersect image boundaries. This can be done over steerable wavelets [15, 16], while preserving the frame constants $A_\mathcal{W}$ and $B_\mathcal{W}$. The resulting wavelet $\psi_\lambda(\cdot - u)$ are supported in Λ_d. They are still indexed by $(\lambda, u) \in \Gamma$. If $\lambda = r_{-\ell}\xi \neq 0$ for $1 \leq j \leq J$, $0 \leq \ell < L$ then $\sum_{u\in\Lambda_d} \psi_\lambda(u) = 0$. If $\lambda = 0$ then $\sum_{u\in\Lambda_d} \psi_0(u) = 2^J$. Since $J \leq (\log_2 d)/2$, there are at most $L(\log_2 d)/2 + 1$ different frequency channels λ. For $\lambda = r_{-\ell}\xi \neq 0$, ψ_λ is translated by $u = 2^{j-1}n \in \Lambda_d$ which yields $2^{-j+1}d$ wavelet coefficients. The total number of wavelets coefficients is about $4Ld/3$ if $J = (\log_2 d)/2$.

The mother wavelet ψ is chosen in order to obtain a sparse wavelet representation of realizations of X, with few large amplitude wavelet coefficients. This sparsity highlights non-Gaussian properties. Figure 2 displays the modulus and phase of wavelet coefficients of the vorticity field of a turbulent flow. This flow is obtained by running the 2D Navier Stokes equation with periodic boundary conditions, initialized with a random Gaussian field [18]. After a fixed time, it defines a stationary but non-Gaussian random process. For each scale and orientation, large amplitude modulus coefficients are located at positions where the image has sharp transitions, and the phase depends upon the position of these sharp transitions.

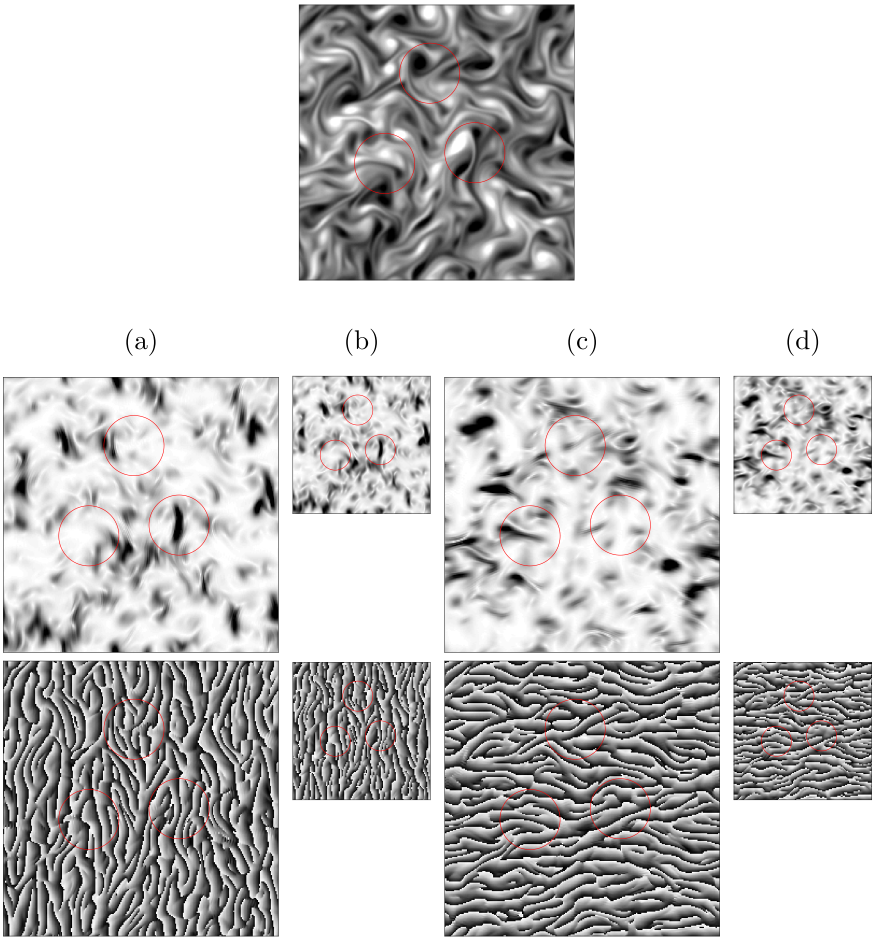

Figure 2: Top: turbulent velocity field. Bottom: Each image gives the modulus (above) or the phase (below) of wavelet coefficients $x \star \psi_\lambda(u)$ for different frequency channels $\lambda = 2^{-j} r_{-\ell} \xi$. Large modulus coefficients are shown in black. The columns correspond to different scales and angles (j, ℓ). (a): $(j, \ell) = (1, 0)$. (b): $(j, \ell) = (2, 0)$. (c): $(j, \ell) = (1, L/4)$. (d): $(j, \ell) = (2, L/4)$.

Wavelet covariance A wavelet transform defines a linear representation $\mathcal{R} = \mathcal{W}$ indexed by $v = (\lambda, u)$. Similarly to Fourier coefficients, we show that wavelet coefficients have a covariance which nearly vanish at different frequencies.

Similarly to Fourier coefficients, wavelet coefficients have a zero mean at non-zero frequencies. If $\lambda \neq 0$ then $\mathbb{E}(X \star \psi_\lambda(u)) = 0$ because $\sum_u \psi_\lambda(u) = 0$. If $\lambda = 0$ then $\sum_u \psi_0(u) = 2^J$ so $\mathbb{E}(X \star \psi_0(u)) = 2^J \mathbb{E}(X(u))$. The covariance at $v = (\lambda, u)$ and $v' = (\lambda', u')$ is

$$K_{\mathcal{W}}(v, v') = \mathrm{Cov}(X \star \psi_\lambda(u)\,,\, X \star \psi_{\lambda'}(u'))\ .$$

It depends on $u - u'$ because X is stationary. Let $\widehat{\psi}_\lambda(\omega)$ be the discrete Fourier transform of $\psi_\lambda(u)$ defined in (9). Since wavelet coefficients are convolutions, covariance values can be rewritten from the power spectrum $\widehat{K}(\omega) = \frac{1}{d}\mathrm{Cov}(\widehat{X}(\omega), \widehat{X}(\omega))$ of X

$$K_{\mathcal{W}}(v, v') = \sum_{\omega \in \Lambda_d} \widehat{K}(\omega)\, \widehat{\psi}_{\lambda'}(\omega)\, \widehat{\psi}^*_\lambda(\omega)\, e^{i(u-u').\omega}. \tag{29}$$

It results that $K_{\mathcal{W}}(v, v') = 0$ if $\widehat{\psi}_\lambda(\omega)\, \widehat{\psi}_{\lambda'}(\omega) = 0$ for all ω. Since $\widehat{\psi}_\lambda(\omega)$ is non-negligible only if $|\omega - \lambda| \leq C|\lambda|$, the covariance $K_{\mathcal{W}}(v, v')$ is non-negligible for $\lambda \neq \lambda'$ only if

$$\frac{|\lambda - \lambda'|}{|\lambda| + |\lambda'|} \leq C\,. \tag{30}$$

It shows that similarly to Fourier coefficients, wavelet covariances are negligible across frequencies which are sufficiently far apart.

Maximum entropy wavelet graph model A maximum entropy model conditioned by wavelet covariances is Gaussian because the wavelet transform is linear. Figure 3(a) gives a realization of a stationary turbulent flow X. A low-dimensional maximum entropy model is defined on a graph of covariance coefficients $K_{\mathcal{W}}(v, v')$ for $v' = (\lambda', u')$ in a small neighborhood of $v = (\lambda, u)$. Since wavelet coefficients are nearly decorrelated across frequencies, at each scale 2^j, the neighborhood of $v = (\lambda, u)$ is defined as the set of $v' = (u', \lambda')$ such that $\lambda' = \lambda$ and $|u - u'| \leq 2^{j-1}\Delta$ for a fixed Δ. Covariances are thus specified over a spatial range proportional to the scale. This is a foveal neighborhood which is sufficient to approximate the covariances of large classes of random processes such as fractional Brownian motions [19]. Since $(u, u') = 2^{j-1}(n, n')$, the neighborhoods of all v have the same size, which is smaller than $(2\Delta + 1)^2$.

Since X is stationary, its probability distribution is invariant to the group G of translations. The number of covariance coefficients that must be estimated is equal to the number $|E_G|$ of edges in the graph modulo translations. It is equal to the number of wavelet frequencies λ multiplied by the size of each neighborhood, and thus bounded by $(JL+1)(2\Delta+1)^2 = O(\log_2 d)$. This model size is much smaller than the image size d when d is large.

Figure 3(b) shows a realization of the maximum entropy Gaussian model $\widetilde{X}$. It is conditioned by wavelet covariances on a foveal graph with $\Delta = 2$. The wavelet transform is computed with a bump wavelet specified in Appendix D, with $J = 5$, $L = 16$ and $d = 256^2$. In this case $|E_G|/d = 3.6\,10^{-2}$. The covariances are estimated

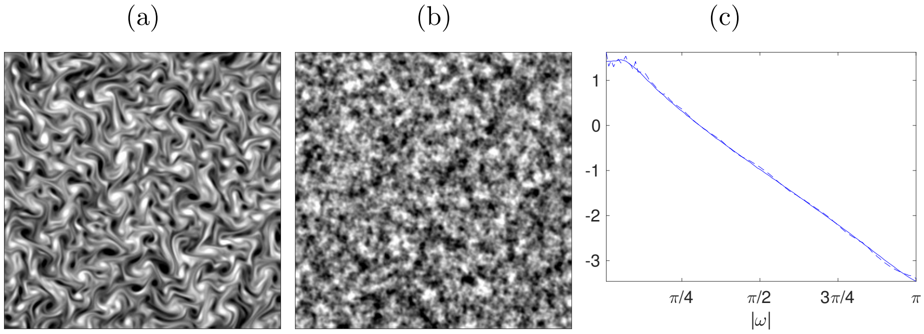

Figure 3: (a): Realization of a stationary turbulent vorticity field X. (b): Realization of a Gaussian maximum entropy model $\widetilde{X}$ calculated from wavelet covariances estimated on a foveal graph. (c): The full line and dashed lines are the logarithms of the power spectrum of X and of the power spectrum of $\widetilde{X}$ respectively as a function of the radial frequency $|\omega|$.

from a single realization of X. The calculation of the Lagrange multipliers in (2) is explained in Appendix A. To measure the accuracy of this model we compare the power spectrum of X and the Gaussian model $\widetilde{X}$. Both processes are isotropic so Figure 3(c) gives the radial log power spectrum of X and $\widetilde{X}$ as a function of $|\omega|$. These power spectrum are nearly the same, which means that the wavelet covariance graph gives an accurate estimation of the second order moments of X from a single realization.

The realization of the Gaussian model in Figure 3(b) has a geometry which is very different from the turbulence flow X in Figure 3(a). It shows that X is highly non-Gaussian, which also appears in Figure 2. The wavelet coefficients of a stationary process X are not correlated at different scales and angles, which would imply that they are independent if X was Gaussian. On the contrary, Figure 2 shows that the the modulus and phases of wavelet coefficients of X are strongly dependent across scales and angles. High amplitude modulus coefficients are located in the same spatial neighborhoods because they are produced by the same sharp transitions of the flow. Next section explains how to capture this dependence with phase harmonics.

3.2 Wavelet Phase Harmonic Covariance

As in the Fourier case, phase harmonics create correlations between wavelet coefficients across different frequency bands. We study the properties of the resulting covariance matrix, and the role of sparsity.

Wavelet phase harmonics To specify the dependence across frequencies, we apply a phase harmonic operator to wavelet coefficients:

$$\widehat{\mathcal{H}}(\mathcal{W}x) = \left\{ \widehat{h}(k) \, [x \star \psi_\lambda(u)]^k \right\}_{(\lambda,u)\in\Gamma, k\in\mathbb{Z}} .$$

The coefficients of $\mathcal{R} = \widehat{\mathcal{H}}\mathcal{W}$ are indexed by $v = (\lambda, k, u)$. The covariance coefficients of $\widehat{\mathcal{H}}(\mathcal{W}X)$ are

$$K_{\widehat{\mathcal{H}}\mathcal{W}}(v, v') = \widehat{h}(k)\,\widehat{h}(k')^*\,\mathrm{Cov}([X \star \psi_\lambda(u)]^k\,,\,[X \star \psi_{\lambda'}(u')]^{k'}).$$

Since X is stationary, it only depends on $u - u'$. Wavelet harmonic covariances only need to be calculated for $k \geq 0$ and $k' \geq 0$. Indeed, since X is real and $\psi(-u) = \psi^*(u)$ one can verify that $K_{\widehat{\mathcal{H}}\mathcal{W}}(v, v')$ does not change its value if (λ, k) becomes $(-\lambda, -k)$ or if (k, k') becomes $(-k, -k')$. Such wavelet harmonic covariances have first been computed by Portilla and Simoncelli [8] to characterize the statistics of image textures. Their representation correspond to (k, k') equal to $(0, 0)$, $(1, 1)$, $(1, 2)$, which amounts to choosing $\hat{h}(k) = 1_{[0,2]}(k)$.

Since $\widehat{\mathcal{H}}$ and $\mathcal{W}$ are bi-Lipschitz the operator $\widehat{\mathcal{H}}\mathcal{W}$ is also bi-Lipschitz, with lower and upper bounds $A_{\mathcal{H}}\,A_{\mathcal{W}}$ and $B_{\mathcal{H}}\,B_{\mathcal{W}}$. Proposition 2.1 implies that

$$A_{\mathcal{H}}\,A_{\mathcal{W}}\,\sigma^2(X) \leq \sigma^2(\widehat{\mathcal{H}}(\mathcal{W}X)) \leq B_{\mathcal{H}}\,B_{\mathcal{W}}\,\sigma^2(X), \tag{31}$$

which controls the variance of wavelet harmonic coefficients.

Rectified neural network coefficients Ustyuzhaninov et. al. in [9] have shown that one can get good texture synthesis from the covariance of a one-layer convolutional neural network, computed with a rectifier. In the following we show that these statistics are equivalent to phase harmonic covariances, computed with a rectifier phase window $h(\alpha)$.

Section 2.3 proves that $\widehat{\mathcal{H}} = \mathcal{F}_\alpha \mathcal{H}$, where $\mathcal{H}$ computes a phase windowing of wavelet coefficients

$$\mathcal{H}(\mathcal{W}x) = \Big\{|x \star \psi_\lambda(u)|\,h(\varphi(x \star \psi_\lambda(u)) + \alpha)\big\}_{(\lambda,u)\in\Gamma,\alpha\in[0,2\pi]}.$$

The covariance of $\widehat{\mathcal{H}}(\mathcal{W}X)$ and $\mathcal{H}(\mathcal{W}X)$ thus satisfy $K_{\widehat{\mathcal{H}}\mathcal{W}} = \mathcal{F}_\alpha\,K_{\mathcal{H}\mathcal{W}}\,\mathcal{F}_\alpha^{-1}$. The following proposition proves that $K_{\mathcal{H}\mathcal{W}}$ gives the covariance of rectified wavelet coefficients if h is a rectifier phase window.

Proposition 3.1. *Let $v = (\lambda, \alpha, u)$ and $v' = (\lambda', \alpha', u')$. For a rectifier phase window $h(\alpha) = \rho(\cos\alpha)$*

$$K_{\mathcal{H}\mathcal{W}}(v, v') = \mathrm{Cov}\Big(\rho(X \star \psi_{\lambda,\alpha}(u))\,,\,\rho(X \star \psi_{\lambda',\alpha'}(u'))\Big) \tag{32}$$

with $\psi_{\lambda,\alpha}(u) = \mathrm{Real}(e^{-i\alpha}\psi_\lambda(u))$.

Proof: We proved in (17) that if $h(\alpha) = \rho(\cos\alpha)$ then

$$\mathcal{H}(z) = \{\rho(\mathrm{Real}(e^{i\alpha}z))\}_{\alpha\in[0,2\pi]}.$$

It results that

$$K_{\mathcal{H}\mathcal{W}}(v, v') = \mathrm{Cov}(\rho(\mathrm{Real}(e^{i\alpha}X \star \psi_\lambda(u))), \rho(\mathrm{Real}(e^{i\alpha'}X \star \psi_{\lambda'}(u')))\rangle,$$

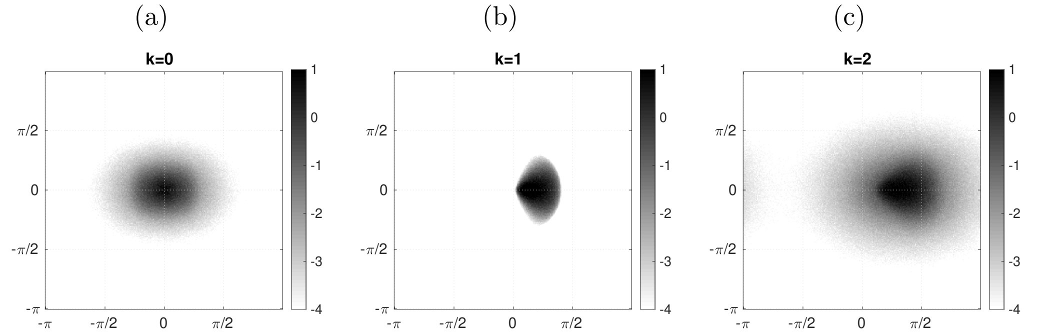

Figure 4: Power spectrum of $[X \star \psi_\lambda(u)]^k$ for a fixed λ, for the turbulent vorticity flow at the top of Figure 5(a), and different k. The power spectrum is shown in log base 10. (a): $k = 0$. (b): $k = 1$. (c): $k = 2$.

which proves (32). □

Rectified wavelet coefficients $\rho(X \star \psi_{\lambda,\alpha}(u))$ can be interpreted as one-layer convolutional network coefficients, computed with wavelet filters $\psi_{\lambda,\alpha}$ of different frequencies λ and phases α. The difference with the statistics used by Ustyuzhaninov et. al. in [9] relies in the choice of network filters. They use local cosine or random filters in their network as opposed to steerable wavelets.

Sparse harmonic covariance Proposition 2.2 specifies the properties of a Fourier phase harmonic covariance. We qualitatively explain, without proof, why $\mathrm{Cov}([X \star \psi_\lambda(u)]^k, [X \star \psi_{\lambda'}(u')]^{k'})$ has similar sparsity properties.

For a fixed λ and k, $[X \star \psi_\lambda(u)]^k$ is a stationary random vector in u. The covariance of $[X \star \psi_\lambda(u)]^k$ and $[X \star \psi_{\lambda'}(u')]^{k'}$ is non-zero only if their power-spectrum have a support which overlap. We give a necessary condition by approximating these spectrum supports.

For $k = 1$, the spectrum of $X \star \psi_\lambda(u)$ has an energy concentrated at frequencies where the Fourier transform of $\widehat{\psi}_\lambda$ is concentrated, which is included in a ball centered at λ of radius $C|\lambda|$. For $|k| > 1$, [4] explains that the Fourier transform of $[X \star \psi_\lambda(u)]^k$ and hence its power spectrum is concentrated in a ball centered at $k\lambda$ of radius $|k|C|\lambda|$. If $k = 0$ then the spectrum of $|X \star \psi_\lambda(u)|$ is concentrated in a ball centered at the 0 frequency, of radius $C|\lambda|$. This is illustrated numerically in Figure 4 which displays the power spectrum of $[X \star \psi_\lambda(u)]^k$ for $k = 0, 1, 2$. In this case, X is a stationary vorticity field shown at the top of Figure 3(a). These power spectrum are estimated for a fixed λ from 100 independent realizations of X. For $k = 1$, the power spectrum is supported over the Fourier support of $\widehat{\psi}_\lambda$. For $k = 2$ its support is approximately dilated by 2, whereas for $k = 0$ it is centered at the zero frequency.

It results that the spectrum of $[X \star \psi_\lambda(u)]^k$ and $[X \star \psi_{\lambda'}(u')]^{k'}$ have a support which overlap only if

$$|k\lambda - k'\lambda'| \leq C(\max(|k|, 1)\,|\lambda| + \max(|k'|, 1)\,|\lambda'|)\ . \tag{33}$$

Similarly to the Fourier case, if $k = k' = 0$ then this condition is satisfied for any pair of frequencies (λ, λ'). It corresponds to covariances of modulus coefficients $\mathrm{Cov}(|X \star \psi_\lambda(u)|\,,\,|X \star \psi_{\lambda'}(u')|)$, which are typically non-zero when the process is non-Gaussian.

If $k \neq 0$ and $k' \neq 0$ then non-negligible values of $\mathrm{Cov}([X \star \psi_\lambda(u)]^k, [X \star \psi_{\lambda'}(u')]^{k'})$ occur for $k\lambda \approx k'\lambda'$. This is similar to the vanishing property (22) of Fourier harmonic coefficients, at frequencies which are colinear. Since $\lambda = 2^{-j} r_{-\ell}\xi$ and $\lambda' = 2^{-j'} r_{-\ell'}\xi$, it requires that $2^{j-j'} \approx |k'|/|k|$ and that $\ell \approx \ell'$.

Diagonal covariance coefficients have similar properties as Fourier coefficients. They are specified by $\mathrm{Cov}(|X \star \psi_\lambda(u)|, |X \star \psi_\lambda(u)|)$ and $\mathrm{Cov}(X \star \psi_\lambda(u), X \star \psi_\lambda(u))$, which do not depend upon u. Moreover when $\lambda \neq 0$,

$$\frac{\mathrm{Cov}(|X \star \psi_\lambda(u)|, |X \star \psi_\lambda(u)|)}{\mathrm{Cov}(X \star \psi_\lambda(u), X \star \psi_\lambda(u))} = 1 - \frac{\mathbb{E}(|X \star \psi_\lambda(u)|)^2}{\mathbb{E}(|X \star \psi_\lambda(u)|^2)}.$$

The ratio $\mathbb{E}(|X \star \psi_\lambda(u)|)^2/\mathbb{E}(|X \star \psi_\lambda(u)|^2)$ measures the sparsity of wavelet coefficients $X \star \psi_\lambda(u)$. If X is Gaussian then $X \star \psi_\lambda(u)$ is a complex Gaussian random variable and this ratio is $\frac{\pi}{4}$ for all λ. A ratio smaller than $\frac{\pi}{4}$ implies that wavelet coefficients $X \star \psi_\lambda(u)$ are sparse and hence that X is non-Gaussian.

To increase the accuracy of a maximum entropy model $\widetilde{X}$, the Kullback-Leibler divergence (8) shows that we must minimize its entropy. As in the Fourier case, an upper bound of the entropy is obtained as a sum of the entropy of the marginals of wavelet coefficients. The marginal entropies get smaller by reducing the sparsity ratio $\mathbb{E}(|X \star \psi_\lambda(u)|)^2/\mathbb{E}(|X \star \psi_\lambda(u)|^2)$. To minimize this upper bound of the model entropy, this suggests by finding a mother wavelet ψ which yields sparse coefficients.

Gaussianity test Even though ψ_λ and $\psi_{\lambda'}$ may have disjoint Fourier supports, the previous analysis showed that one can find k and k' such that the spectrum of $[X \star \psi_\lambda(u)]^k$ and $[X \star \psi_{\lambda'}(u')]^{k'}$ overlap, for example with $k = k' = 0$. A priori, the covariance $K_{\widehat{H}\mathcal{W}}(v, v')$ is then non-zero, unless X is Gaussian in which case these coefficients vanish, as proved by the following theorem.

Proposition 3.2. *Let (λ, λ') be such that $\widehat{\psi}_\lambda\, \widehat{\psi}_{\lambda'} = 0$. If X is Gaussian and stationary then $K_{\widehat{H}\mathcal{W}}(v, v') = 0$ if $v = (\lambda, k, u)$ and $v' = (\lambda', k', u')$, for any (k, u, k', u').*

Proof: If X is Gaussian then $X \star \psi_\lambda(u)$ and $X \star \psi_{\lambda'}(u')$ are jointly Gaussian random variables. Equation (29) proves that if $\widehat{\psi}_\lambda\, \widehat{\psi}_{\lambda'} = 0$ then $\mathrm{Cov}(X \star \psi_\lambda(u), X \star \psi_{\lambda'}(u')) = 0$. These Gaussian random variables are uncorrelated and therefore independent. It results that $[X \star \psi_\lambda(u)]^k$ and $[X \star \psi_{\lambda'}(u')]^{k'}$ are also independent and thus have a covariance which is zero.□

This proposition gives a test of Gaussianity by considering two frequencies λ and λ' where $|\lambda - \lambda'|$ is sufficiently large so that the support of $\widehat{\psi}_\lambda$ and $\widehat{\psi}_{\lambda'}$ do not overlap. If $k\lambda \approx k'\lambda'$ then wavelet harmonic covariances are zero if X is Gaussian but it is typically non-zero if X is not Gaussian.

4 Microcanonical Models

Section 2.1 introduces macrocanonical maximum entropy models conditioned by covariance coefficients of a representation $\mathcal{R}(X)$. The resulting maximum entropy distribution $\tilde{p}$ depend upon Lagrange coefficients which are computationally very expensive to calculate, despite the development of efficient algorithms [20]. Section 4.1 reviews maximum entropy microcanonical models which avoid computing these Lagrange multipliers. Section 4.2 gives an alternative microcanonical model which is calculated with a faster algorithm, which transports a maximum entropy measure by gradient descent.

4.1 Maximum Entropy Microcanonical Models

Microcanonical models avoid the calculation of Lagrange multipliers and are guaranteed to exist as opposed to macrocanonical models. We first specify covariance estimators and then briefly review the properties of these microcanonical models.

Microcanonical models rely on an ergodicity property which insures that covariance estimations concentrate near the true covariance $K_{\mathcal{R}}$ when d is sufficiently large. Let G be a known group of linear unitary symmetries of the density p of X. It includes translations because X is stationary. An estimation of the covariance $K_{\mathcal{R}}$ over the edge set E_G is computed in (4) from a single realization $\bar{x}$ of X, by transforming $\bar{x}$ with all $g \in G$. To differentiate realizations of X from other $x \in R^d$, we associate a similar covariance estimation to any $x \in \mathbb{R}^d$. This covariance is centered on the empirical mean $\widetilde{M}_{\mathcal{R}}$ computed in (3) and defined by

$$\widetilde{K}_{\mathcal{R}x}(v, v') = \frac{1}{|G|} \sum_{g \in G} \Big(\mathcal{R}_v(g.x) - \widetilde{M}_{\mathcal{R}}(v)\Big) \Big(\mathcal{R}_{v'}(g.x) - \widetilde{M}_{\mathcal{R}}(v')\Big)^* . \qquad (34)$$

It is invariant to the action of any $g \in G$ on x.

We denote $\|K_{\mathcal{R}}\|^2_{E_G} = \sum_{(v,v') \in E_G} |K_{\mathcal{R}}(v, v')|^2$. We shall suppose that X satisfies the following covariance ergodicity property over E_G:

$$\forall \epsilon > 0, \quad \lim_{d \to \infty} \text{Prob}(\|\widetilde{K}_{\mathcal{R}X} - K_{\mathcal{R}}\|_{E_G} \leq \epsilon) = 1. \qquad (35)$$

To satisfy this property the number $|E_G|$ of covariance moments must be small compared to the dimension d of X. The ergodicity hypothesis implies that when d is sufficiently large, with high probability, the empirical covariance $\widetilde{K}_{\mathcal{R}\bar{x}}$ of a realization $\bar{x}$ of X is close to the true covariance $K_{\mathcal{R}}$ over E_G.

Given a realization $\bar{x}$ of X, a microcanonical set of width ϵ is the set of all $x \in \mathbb{R}^d$ which have nearly the same covariance estimations as $\bar{x}$:

$$\Omega_\epsilon = \{x \in \mathbb{R}^d : \|\widetilde{K}_{\mathcal{R}x} - \widetilde{K}_{\mathcal{R}\bar{x}}\|_{E_G} \leq \epsilon\}. \qquad (36)$$

A maximum entropy microcanonical model has a probability distribution of maximum entropy supported in Ω_ϵ. The set Ω_ϵ is bounded. Indeed, if $x \in \Omega_\epsilon$ then $\|\widetilde{K}_{\mathcal{R}x}\|_{E_G} \leq \|\widetilde{K}_{\mathcal{R}\bar{x}}\|_{E_G} + \epsilon$. Since E_G includes all diagonal coefficients (v, v) for $v \in V$, one can derive an upper bound for $\|\mathcal{R}x\|$. Since $\mathcal{R}$ is bi-Lipschitz, we also obtain an upper bound for $\|x\|$. Since Ω_ϵ is compact, the maximum entropy distribution is uniform in Ω_ϵ relatively to the Lebesgue measure. A major issue in statistical physics is to find sufficient conditions to prove the Boltzmann equivalence principle which guarantees the convergence of microcanonical and macrocanonical models towards the same Gibbs measures when d goes to ∞ [21]. This involves the proof of a large deviation principle which expresses concentration properties of the covariance of $\mathcal{R}(X)$. If $\mathcal{R}$ is continuous and bounded so that the interaction potential $(\mathcal{R} - M_{\mathcal{R}})(\mathcal{R} - M_{\mathcal{R}})^*$ is also continuous and bounded, and if there is no phase transition, which means that the limit is a unique Gibbs measure, then one can prove that microcanonical and macrocanonical measures converge to the same limit for an appropriate topology [22, 23]. The bounded hypothesis is not necessary and may not be satisfied.

The ergodicity property (35) guarantees that X concentrates in Ω_ϵ with a high probability when d is sufficiently large. The microcanonical set Ω_ϵ may however be much larger than the set where X concentrates, which means that the maximum entropy microcanonical distribution may have a much larger entropy than the entropy of X. As in the macrocanonical case, the representation $\mathcal{R}$ must be optimized in order to reduce the maximum entropy, which motives the use of sparse representations.

4.2 Gradient-Descent Microcanonical Models

Sampling a maximum entropy microcanonical set requires to use Monte Carlo algorithms. They are computationally very expensive when the number of moments $|E_G|$ and the dimension d are large, because their mixing time become prohibitive [24]. Following the approach in [11], we approximate these microcanonical models with a gradient descent algorithm.

Gradient-descent microcanonical models are computed by transporting an initial Gaussian white noise measure into the microcanonical set Ω_ϵ. This transport is calculated with a gradient descent which progressively minimizes

$$f(x) = \|\widetilde{K}_{\mathcal{R}x} - \widetilde{K}_{\mathcal{R}\bar{x}}\|^2_{E_G}, \tag{37}$$

with a sufficiently large number of iterations so that $f(x) < \epsilon$ and hence $x \in \Omega_\epsilon$.

The initial Gaussian white noise is a maximum entropy distribution conditioned by a variance σ^2. This variance is chosen to be an upper bound of the empirical variance of $\bar{x}(u)$ along u, calculated from diagonal coefficients of $\widetilde{K}_{\mathcal{R}\bar{x}}$. It guarantees that the resulting white noise has an entropy larger than the microcanonical model. The gradient descent progressively reduces this entropy while transporting the measure towards Ω_ϵ. Entropy reduction bounds are computed in [11].

The initial x_0 is Gaussian white noise. For all $t \geq 0$, the gradient descent iteratively computes

$$x_{t+1} = x_t - \eta \, \nabla f(x_t) \ .$$

This operation transports the probability measure μ_t of x_t into a measure μ_{t+1} of x_{t+1}. We stop the algorithm at a time $t = T$ which is large enough so that $f(x_T) < \epsilon$ with a high probability. Since $f(x)$ is not convex, there is no guarantee that $f(x_T) < \epsilon$ even for large T. In numerical calculations in Section 5, we may use several initializations, typically 10, to evaluate the model.

The empirical covariance $\widetilde{K}_{\mathcal{R}x}$ is computed in (34) with an average over all symmetries of a known group G of linear unitary operators in $\mathbb{R}^d$. The following theorem proves that G is also a group of symmetries of the probability measures obtained by gradient descent.

Theorem 4.1. *For any $t \geq 0$, the probability measure μ_t of x_t is invariant to the action of G.*

The proof is in Appendix E. It is a minor modification of the proof in [11]. We replace the gradient descent method by L-BFGS algorithm with line search [25]. It has a faster and more accurate numerical convergence. Since G includes translations, this theorem implies that each x_t is stationary. It is shown in [11] that the transported density μ_T supported in Ω_ϵ may be different from the maximum entropy density in Ω_ϵ because the gradient descent reduces too much the entropy. In general, the precision of these microcanonical gradient descent models are not well understood. However, this theorem proves that the gradient descent preserves all known symmetries of the distribution of X, which is also true for a maximum entropy measure,

The gradient descent may converge faster by preconditioning $f(x)$. This is done by replacing estimated covariances $\widetilde{K}_{\mathcal{R}x}(v, v')$ by normalized correlation coefficients

$$\frac{\widetilde{K}_{\mathcal{R}x}(v, v')}{\widetilde{K}_{\mathcal{R}\bar{x}}(v, v)^{1/2}\, \widetilde{K}_{\mathcal{R}\bar{x}}(v', v')^{1/2}} \ . \tag{38}$$

Any $g \in G$ is a symmetry of $\widetilde{K}_{\mathcal{R}x}$ and is therefore a symmetry of normalized correlation coefficients, so Theorem 4.1 remains valid with this preconditioning.

5 Foveal Wavelet Harmonic Covariance Models

We study microcanonical models conditioned by wavelet harmonic covariances, computed with different symmetry groups G, and different sufficient statistics set E_G. Section 5.2 introduces foveal models which limit the multiscale spatial range of coefficients in E_G. Section 5.3 gives a methodology to evaluate the precision of different foveal models, with numerical results.

5.1 Rotation and Reflection Symmetries

The covariance $K_{\mathcal{R}}$ is estimated from a single realization of X, by taking advantage of a known group of symmetries G of the the probability distribution of X. For a wavelet phase harmonic representation $\mathcal{R} = \widehat{\mathcal{H}}\mathcal{W}$, we specify the properties of $K_{\mathcal{R}}$ when G includes sign changes, reflections and rotations.

The following proposition proves that some covariance coefficients vanish when symmetries include a sign change or a central reflection. These covariances thus do not need to be included in the sufficient statistics set E_G. The proof is in Appendix F.

Proposition 5.1. *Let $v = (\lambda, k, u)$ and $v' = (\lambda', k', u')$.*
(i) If the sign change $g.x = -x$ is a symmetry of the probability distribution of X then $K_{\widehat{H}\mathcal{W}}(v, v') = \widetilde{K}_{\widehat{H}\mathcal{W}x}(v, v') = 0$ if $k + k'$ is odd.
(ii) If the central reflection $g.x(u) = x(-u)$ is a symmetry of the probability distribution of X and $\widehat{h}$ is real then $K_{\widehat{H}\mathcal{W}}(v, v')$, and $\widetilde{K}_{\widehat{H}\mathcal{W}x}(v, v')$ are real.

Isotropic models An isotropic random process X has a probability distribution which is invariant by rotations. The group G then includes all translations and rotations. We show that $K_{\widehat{\mathcal{H}}\mathcal{W}}$ becomes sparse after applying a Fourier transform on rotation angles.

If $g = r_\eta$ is a rotation by $2\pi\eta/L$ for $0 \leq \eta < L$ then $(r_\eta.x) \star \psi_\lambda(u) = x \star \psi_{r_\eta\lambda}(r_{-\eta}u)$ where $r_\eta\lambda = 2^{-j} r_{\eta-\ell}\xi$ is the rotation of $\lambda = 2^{-j} r_{-\ell}\xi$. To eliminate the effect of the change of position $r_{-\eta}u$ on covariance coefficients, we only keep wavelet harmonic covariances at a same spatial position. This means that $v' = (\lambda', k', u')$ is a neighbor of $v = (\lambda, k, u)$ in the covariance graph model only if $u = u'$. It implies that $(v, v') \in E_G$ only if v and v' have the same spatial position.

Isotropy is a form of stationarity along rotations angles. To diagonalize angular covariance matrices, we use a discrete Fourier transform along rotations written $\mathcal{F}_\ell$. The discrete Fourier transform of $y(\ell)$ for $0 \leq \ell < L$ at a frequency $0 \leq m < L$ is

$$(\mathcal{F}_\ell y)(m) = \sum_{\ell=0}^{L-1} y(\ell)\, e^{-i2m\pi\ell/L}.$$

The representation $\mathcal{R}(X) = \mathcal{F}_\ell\, \widehat{\mathcal{H}}(\mathcal{W}X)$ computes $\mathcal{F}_\ell([X \star \psi_\lambda(u)]^k)$ for $\lambda = 2^{-j} r_{-\ell}\xi$ with (j, u, k) fixed and ℓ varying. It is indexed by $v = (j, m, k, u)$. The covariance matrix of $\mathcal{R}(X)$ is

$$K_{\mathcal{R}} = \mathcal{F}_\ell K_{\widehat{\mathcal{H}}\mathcal{W}}\, \mathcal{F}_\ell^{-1}.$$

We consider the restriction of $K_{\mathcal{R}}$ to E_G. The next theorem proves that if X is isotropic then $K_{\mathcal{R}}$ has diagonal angular Fourier matrices. Isotropic processes X may also have a probability distribution which are invariant to line reflections. A line reflection of orientation η computes $g.x(u) = x(u_\eta)$, where u_η is symmetric to u relatively to a line going through the origin in $\mathbb{R}^2$, with an orientation $\eta \in [0, 2\pi]$. If X is isotropic and invariant to a line reflection for an angle η then it is invariant to line reflections for any $\eta \in [0, 2\pi]$. The following theorem applies to the bump steerable wavelets used in numerical experiments.

Theorem 5.1. *Let $v = (j, m, k, u)$, $v' = (j', m', k', u)$ and $u = (u_1, u_2)$. If the probability distribution of X stationary and isotropic and $\mathcal{R} = \mathcal{F}_\ell\, \widehat{\mathcal{H}}\mathcal{W}$ then*

$$K_{\mathcal{R}}(v, v') = \widetilde{K}_{\mathcal{R}x}(v, v') = 0 \quad \text{if } m \neq m'. \tag{39}$$

Furthermore, if the distribution of X is invariant to line reflections and if the wavelet satisfies $\psi(u_1, -u_2) = \psi(u_1, u_2)$ and $\phi(u_1, -u_2) = \phi(u_1, u_2)$ then $K_{\mathcal{R}}(v, v')$ and $\widetilde{K}_{\mathcal{R}x}(v, v')$ are real if $m = m'$.

The proof is in Appendix G. This theorem proves that the sufficient statistics set can be reduced to diagonal angular coefficients $(v, v') \in E_G$ with $m = m'$. It reduces its size by a factor L. Invariance to line reflections implies that it is sufficient to keep the real part of these diagonal values. The rotation invariance strictly applies to the process defined on a continuous variable $u \in \mathbb{R}^2$. On discrete images it is not valid at the finest scale because of the square sampling grid, and it is not valid at the largest scale because of their square support. For preconditioning, the covariance of the isotropic model is also normalized with (38) and the Fourier transform $\mathcal{F}_\ell$ is then applied on the normalized coefficients.

5.2 Foveal Wavelet Harmonic Covariance

A wavelet harmonic covariance model is defined by the choice of the harmonic weights $\hat{h}$, by the symmetry group G and neighborhood relations which defines the edge set E. In the following we shall impose that $\hat{h}(k) = 1_{[k_{\min}, k_{\max}]}(k)$, which limits harmonic exponents in the range $[k_{\min}, k_{\max}]$. We define several foveal models which capture different properties.

Foveal models Wavelet harmonic coefficients are indexed by $v = (\lambda, k, u)$, with $\lambda = 2^{-j} r_{-\ell} \xi$ and $u = 2^{j-1} n$. The covariance graph model is specified by the neighborhoods $\mathcal{N}_v$. A foveal model defines neighborhoods whose size doe not depend upon v. The range of spatial, scale and angular parameters is limited by three parameters Δ_n, Δ_j and Δ_ℓ. A vertex $v' = (\lambda', k', u')$ with $\lambda' = 2^{-j'} r_{-\ell'} \xi$ and $u' = 2^{j'-1} n'$ is a neighbor of $v = (\lambda, k, u)$ only if

$$|n - n'| \leq \Delta_n \ , \ |j - j'| \leq \Delta_j \ , \ |\ell - \ell'| \leq \Delta_\ell \ , \ (k, k') \in [k_{\min}, k_{\max}]^2 \ .$$

A foveal model has a spatial range proportional to the scale. Long range spatialcorrelations are partly captured because $|u - u'| = |n 2^{j-1} - n' 2^{j'-1}|$ become large at large scales. It provides high frequency correlations between close points and low-frequency correlations between far away points. It is similar to a visual fovea [26]. Such foveal models have been used by Portilla and Simoncelli [8] to synthesize image textures.

Because of translation invariance, the sufficient statistics E_G can be defined by setting $n = 0$. Since there are L angles ℓ and at most $(\log_2 d)/2$ scales j, the size of E_G is at most

$$|E_G| = O(L \Delta_\ell (k_{\max} - k_{\min} + 1)^2 (2\Delta_n + 1)^2 \Delta_j \log_2 d).$$

Increasing the values of $k_{\max}, \Delta_j, \Delta_\ell, \Delta_n$ dereases the model bias but it also increases the size $|E_G|$ and hence the variance of the estimation. To ensure the bi-Lipschitz continuity of $\mathcal{R} = \widehat{\mathcal{H}} \mathcal{W}$, we impose that $k_{\min} \leq 1 \leq k_{\max}$. In the following we describe several models of different sizes, which capture different properties of X. Each realization $\bar{x}$ is an image of $d = 256^2$ pixels. We use the bump steerable wavelets of Appendix D, computed on $J = 5$ scales and $L = 16$ angles. The maximum scale 2^J depends on the integral scale of X, which is the distance beyond which all coefficients are nearly independent. A large number of angles L gives a finer angular resolution. We specify 4 models corresponding to different choices of neighborhood parameters. For the first three models, the symmetry group G is reduced to translations where as the last model is also invariant to rotations.

- Model A with $k_{\min} = k_{\max} = 1$. It corresponds to a Gaussian maximum entropy wavelet model of Section 3.1. The covariance of wavelet coefficients is neglected across scales and angles by setting $\Delta_j = 0$, $\Delta_\ell = 0$ and $\Delta_n = 2$. The relative dimension of this model is $|E_G^A|/d = 3.6\,10^{-2}$.

- Model B with $k_{\min} = 0$ and $k_{\max} = 1$. It includes the covariance of the modulus of wavelet coefficients across angles. Covariance across angles at most distant by $\pi/4$ are computed for $k, k' \in \{0, 1\}$ by setting $\Delta_\ell = L/4$. We set $\Delta_j = 0$ and $\Delta_n = 2$, and thus only incorporate covariance across spatial positions. Compared to Model A, spatial correlations are included for $k = k' = 1$ but also for $k = k' = 0$. The relative model size is $|E_G^B|/d = 1.1\,10^{-1}$.

- Model C with $k_{\min} = 0$ and $k_{\max} = 2$. It incorporates covariances of the modulus and phase of wavelet coefficients across scales and angles. Neighborhoods are limited by $\Delta_j = 1$, $\Delta_\ell = L/4$ and $\Delta_n = 2$. Compared to Model B, it incorporates $j' = j + 1$ to capture scale interactions. The phases of wavelet coefficients at a scale 2^j and $2^{j'} = 2^{j+1}$ are correlated with the harmonic exponents $(k, k') = (1, 2)$. The set of (k, k') are restricted to $k' = 0, 1, 2$ when $k = 0$, and $k' = 1, 2$ when $k = 1$. We use the same spatial correlation range as Model B. The relative model size is $|E_G^C|/d = 1.7\,10^{-1}$.

- Model D with $k_{\min} = 0$ and $k_{\max} = 2$ is a rotation invariant version of Model C, with the same Δ_j, Δ_ℓ and (k, k'). This model sets $\Delta_n = 0$ and thus does not capture spatial correlations explicitly. The symmetry group G is composed of translations and L rotations by $2\pi\eta/L$. This rotation invariance is represented by computing a Fourier transform along angles and setting to zero the covariance coefficients according to Theorem 5.1. The model size is therefore much smaller with $|E_G^D|/d = 1.2\,10^{-2}$. Theorem 4.1 proves that the resulting gradient descent microcanonical distribution is invariant to these L rotations.

Visual evaluation The quality of different microcanonical models is first evaluated visually. Among 10 synthesis of each model, we retain the one which yields the smallest loss in (37). The top row of Figure 5 shows a realization $\bar{x}$ of different stationary processes X. The first and second columns display isotropic and non-istotropic vorticity fields of two-dimensional turbulent flows, with a factor 2 zoom on the central part of each image. The third and fourth columns show an image of bubbles and a realization of Multifractal Random Walk [27]. An MRW is a self-similar random process with long range dependencies. We model the increments of MRW, which is a stationary process and limit its maximum correlation scale to 2^5. The next rows give realizations of the microcanonical models A, B, C and D computed from the same realization $\bar{x}$ shown at the top.

As in Figure 3, the foveal Gaussian model A looses most geometric structures. On the contrary the models B, C, D recover most of this geometry. The model B captures

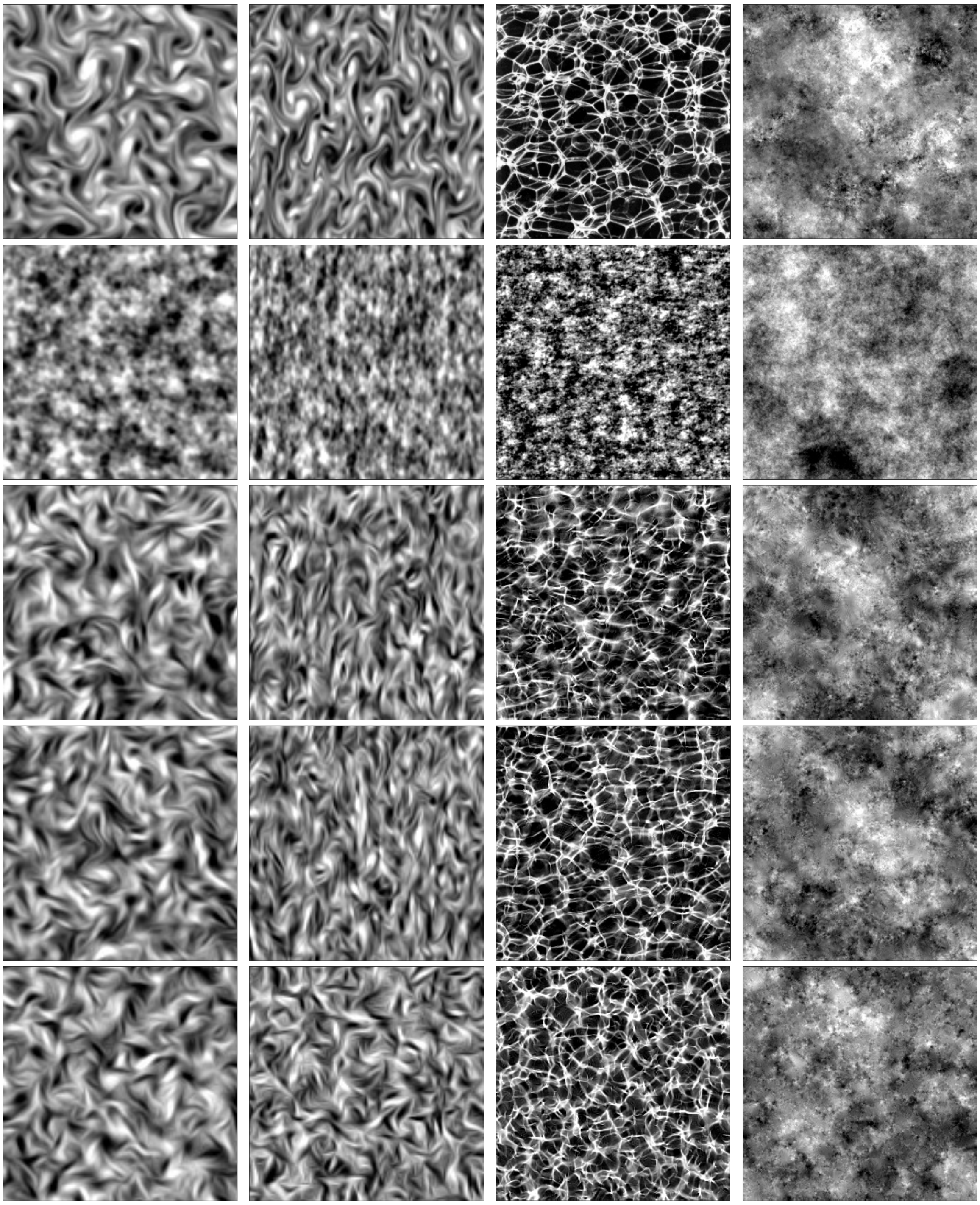

Figure 5: The first row shows a realization $\bar{x}$ of X. The second to fifth row give realizations of microcanonical models A, B, C, D computed from $\bar{x}$. Each column corresponds to a different X. First column: Isotropic turbulent vorticity field. Second column: Non-isotropic turbulent vorticity field. Third column: Bubbles. Fourth column: Multifractal Random Walk.

the correlations of modulus coefficients across angles whereas the model C also impose covariance conditions on phases. There are little visual differences on turbulent flows but it is more apparent on bubble images. Model B does not reproduce closed bubbles whereas model C recovers bubbles having a better geometry. The model D is an isotropic version of model C. Its realizations are therefore isotropic. When X is isotropic then it is visually as precise as C but the variance reduction is not visible. If X is not isotropic, as in the turbulence of the second column, then model D does not reproduce the angular anisotropy.

5.3 Evaluations of Foveal Covariance Models

A microcanonical model $\widetilde{X}$ is conditioned by values of the covariance of $\mathcal{R}_v(X)$ on a graph whose edges E relate neighbor vertices (v, v'). We differentiate two types of model errors. Type I errors result from the use of graph neighborhoods which are too small. In this case, the covariance of $\mathcal{R}(X)$ and $\mathcal{R}(\widetilde{X})$ may be different for $(v, v') \in V^2 - E$. Type II errors are due to the choice of the representation $\mathcal{R}$. Such errors are evaluated by comparing high order moments of X and $\widetilde{X}$, which are not calculated by $\mathcal{R}$. These comparisons are performed for $\mathcal{R} = \widehat{\mathcal{H}\mathcal{W}}$.

Wavelet harmonic covariance error Comparisons are performed over correlation matrices, which normalizes covariance values. Type I errors are calculated by comparing the correlation values of X and $\widetilde{X}$ over a neighborhood which is much larger than the ones used by the different models. This paragraph considers the case where this neighborhood includes all scales and angles, but over a limited spatial range.

We compute correlation coefficients (38) by normalizing $K_\mathcal{R}$ with its diagonal:

$$C_\mathcal{R} = D_\mathcal{R}^{-1/2}\, K_\mathcal{R}\, D_\mathcal{R}^{-1/2} \quad \text{where} \quad D_\mathcal{R} = \mathrm{diag}(K_\mathcal{R}). \tag{40}$$

We estimate $K_\mathcal{R}$ with an empirical average over 100 realizations of X. This gives an accurate estimation of both $D_\mathcal{R}$ and $C_\mathcal{R}$. An empirical estimator $\widetilde{C}_{\mathcal{R}\bar{x}}$ of $C_\mathcal{R}$ is computed from a single realization $\bar{x}$ of X, with the same normalization

$$\widetilde{C}_{\mathcal{R}\bar{x}} = D_\mathcal{R}^{-1/2}\, \widetilde{K}_{\mathcal{R}\bar{x}}\, D_\mathcal{R}^{-1/2}.$$

It is calculated on V_0^2 where $V_0 \subset V$ is a foveal subset of all vertices. It incorporates correlations between all scales and angles, across a limited spatial range.

Let $\|C_\mathcal{R}\|_{op,V_0^2}$ be the operator norm and hence the largest eigenvalue of the symmetric matrix $C_\mathcal{R}$ restricted to V_0^2. The following empirical error measures the estimator error of $C_\mathcal{R}$ from one realization $\bar{x}$

$$\epsilon_{emp} = \frac{\|C_\mathcal{R} - \widetilde{C}_{\mathcal{R}\bar{x}}\|_{op,V_0}}{\|C_\mathcal{R}\|_{op,V_0}}. \tag{41}$$

It is a variance term due to variabilities of realizations $\bar{x}$ of X.

This empirical error is compared to the estimation error of $C_\mathcal{R}$ computed from the different microcanonical models. Let $K_{\mathcal{R}\bar{x}}^{model}$ be the covariance matrix of wavelet harmonic coefficients of a microcanonical model $\tilde{X}$ instead of X. This matrix is estimated

	Isotropic	Non Isotropic
ϵ_{emp}	0.58 (0.05)	0.68 (0.09)
ϵ_A	0.82 (0.01)	0.80 (0.01)
ϵ_B	0.29 (0.02)	0.30 (0.02)
ϵ_C	0.24 (0.02)	0.25 (0.03)
ϵ_D	0.25 (0.02)	1.6 (0.09)

Table 1: Covariance errors (42) models A,B,C,D compared to the empirical error (41), for the isotropic and non-isotropic vorticity fields.

from 10 realizations of $\tilde{X}$. We compare $K^{model}_{\mathcal{R}\bar{x}}$ with $\widetilde{K}_{\mathcal{R}\bar{x}}$ by using the same normalization. We define $C^{model}_{\mathcal{R}\bar{x}}$ by normalizing the $K^{model}_{\mathcal{R}\bar{x}}$ with $D_{\mathcal{R}}$ as in (40). The relative error of the model is defined by

$$\epsilon_{model} = \frac{\|C_{\mathcal{R}} - C^{model}_{\mathcal{R}\bar{x}}\|_{op,V_0}}{\|C_{\mathcal{R}}\|_{op,V_0}}. \tag{42}$$

This error has a variance term because the microcanonical model depends upon a particular realization $\bar{x}$ of X. It has also a bias term because the model is calculated from covariances over a limited set $E_G \subset V_0^2$. Optimizing E_G is a trade-off between the variance and bias terms. If the sufficient statistic set $E_G \subset V_0^2$ is too small and is unable to reproduce the correlations $C_{\mathcal{R}}$ in $V_0^2 - E_G$ then ϵ_{model} is larger than ϵ_{emp}. If $E_G = V_0^2$ then the variance term dominates and $\epsilon_{model} = \epsilon_{emp}$.

We define V_0 with $k_{min} = 0$, $k_{max} = 4$, a maximum range of scales $\Delta_j = J$ and a maximum range of angles $\Delta_\ell = L/2$, but a small translation range $\Delta_n = 2$. Its size is $|V_0| = 8025$, so errors are evaluated over 8025×8025 correlation values. Table 1 compares the empirical error ϵ_{emp} in (41) with the model error ϵ_{model} in (42), for models A, B, C, D. We report the mean of these estimated errors by averaging the operator norms over 10 realizations $\bar{x}$ of X. The standard deviation is given in brackets. These error values are consistent with visual evaluations.

Table 1 show that the Gaussian model A gives a larger error than ϵ_{emp}, because it captures no dependence across scales and angles. It introduces a large model bias. For models B and C, $\epsilon_{model} < \epsilon_{emp}$. The increase of E_G dramatically reduces the model bias, which clearly appears visually. These models are only conditioned by wavelet harmonic covariances over neighbor scales, but the microcanonical model propagates these constraints across all scales. It provides a good approximations of covariance values between far away scales, evaluated in V_0^2. The error of model C is smaller than model B, suggesting that scale correlations through phase also play an important role. This also appears in the visual quality of synthesized bubble images. The model D has a similar error as model C for the isotropic turbulence although it has a lower variance because estimators are averaged across L angular directions. This indicates that the error is dominated by the model bias term and hence that a better model would be obtained by further increasing E_G. This is verified on model D by increasing $\Delta_j = J$

and $k_{max} = 4$, which yields a smaller covariance error 0.20 (0.02) on the isotropic turbulence. On the contrary, the model D has an error which is much larger than model C on the non-isotropic turbulence, which also appears visually.

Long-range spatial correlations Our microcanonical models are computed with foveal sufficient statistics sets, which are conditioned over relatively small spatial neighborhood at each scale 2^j. Such models can therefore introduce an important error if there exists long range spatial correlations between wavelet harmonic coefficients. This paragraph evaluates such errors.

We compare $C_{\mathcal{R}}(v, v')$ and $C^{model}_{\mathcal{R}\bar{x}}(v, v')$ for $v = (\lambda, k, u)$ and $v' = (\lambda', k', u')$ with $\lambda = \lambda'$, $k = k'$ and for large $|u - u'|$. We estimate an average value $C^{model}_{\mathcal{R}}$ of $C^{model}_{\mathcal{R}\bar{x}}$ over 10 realizations $\bar{x}$. The difference between $C^{model}_{\mathcal{R}}$ and $C_{\mathcal{R}}$ corresponds to a bias error term. For $\lambda = 2^{-j} r_{-\ell}\xi$ we define $\overline{C}_{\mathcal{R}}(k, j, a)$ and $\overline{C}^{model}_{\mathcal{R}}(k, j, a)$ as the maximum values of $C_{\mathcal{R}}(v, v')$ and $C^{model}_{\mathcal{R}}(v, v')$ over all ℓ and all (u, u') at a distance $|u - u'| = 2^j a$, for k and j fixed.

Figure 6 compares the correlation values $\overline{C}_{\mathcal{R}}(k, j, a)$ and $\overline{C}^{model}_{\mathcal{R}}(k, j, a)$ as a function of the normalized distance a, for $k = 0, 1$ and $j = 1, 2, 3$. The notion of short and long range correlations is defined relatively to each scale 2^j through the parameter a. For $k = 1$, it corresponds to correlations of wavelet coefficients, which have a fast decay when the distance a increases. At all scales 2^j the long range correlations of wavelet coefficients for models A and D are close to the correlations obtained with the original turbulence and MRW processes.

For $k = 0$, Figure 6 gives the long range correlation of the modulus of wavelet coefficients. It still has a fast decay for the turbulence field but a slow decay for the MRW. The model D gives a much better approximation of long range spatial correlations of turbulence than model A. However there is a residual error at the finest scale $j = 1$ because these foveal models do not capture long range correlations at fine scales. The error is more dramatic for the MRW process where the modulus of wavelet coefficients are correlated over much longer ranges at fine scales. As expected, this evaluation shows that long range correlations at different scales are not fully captured by foveal models which are constrained on a correlation range of the order of $2^j \Delta_n$ at each scale 2^j. One could incorporate these long range spatial covariances by increasing Δ_n but it would increase considerably the model size which is proportional to $(2\Delta_n + 1)^2$.

High order moments Moments of multiple order q can be measured with structure functions at different scale 2^j [28, 29]. We compute the maximum moment of order q of increments between points (u, u') such that $|u - u'| = 2^{j-1}$

$$S(j, q) = \max_{|u-u'|=2^{j-1}} \mathbb{E}(|X(u) - X(u')|^q).$$

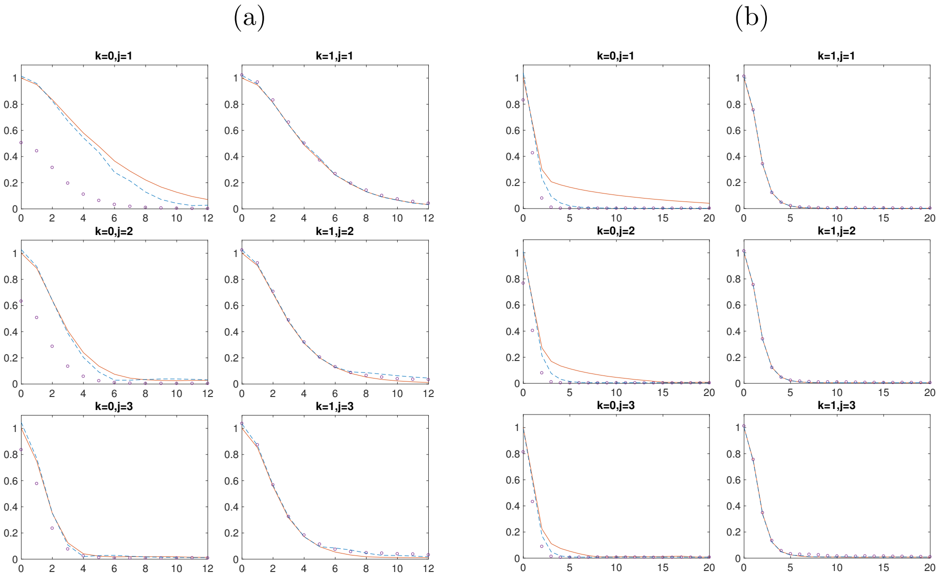

Figure 6: Each graph shows correlation values $\overline{C}_{\mathcal{R}}(k, j, a)$ in full line, as a function of the normalized distance a for a fixed scale 2^j and k. The values of $\overline{C}_{\mathcal{R}}^{model}(k, j, a)$ are shown with circles for model A and dashed lines for model D. The value of k and j is given at the top of each graph. The graphs in (a) are computed for isotropic turbulences and in (b) for increments of a Multifractal Random Walk.

It is compared with the structure function $S_{\mathcal{R}\bar{x}}^{model}(j, q)$ calculated for each microcanonical model $\tilde{X}$, with an error

$$\epsilon_{st}(j, q) = \frac{|S(j, q) - S_{\mathcal{R}\bar{x}}^{model}(j, q)|}{|S(j, q)|}.$$

Table 2 reports the mean and standard deviation (in brackets) of $\epsilon_{st}(j, q)$ for the model A and model D, at two different scales 2^j, for the isotropic turbulence. They are estimated over 10 independent realizations $\bar{x}$. As expected, the errors increase with the order q because foveal covariance model only imposes moments of order $q = 1$ and $q = 2$. The model D creates no significant error on these high-order moments, since the estimated average error is comparable to the estimator standard deviation. It indicates that in this case the covariance terms across scales are sufficient to capture high order moments. On the contrary, the model A has significant errors for $q \neq 2$ because it includes no covariance terms across scales and angles. Table 3 reports the value of these errors for the MRW. The errors of model D remain below the errors of model A but are significant in this case. This is probably due to the fact that the model D does not capture well the long range spatial correlations of the MRW, as shown in Figure 6(b). It introduces an important model error which appears in higher order moments.

This analysis shows that microcanonical models computed from phase harmonic covariances in foveal neighborhoods capture important non-Gaussian properties. It gives accurate models of large classes of processes such as the turbulence examples.

q	1	2	3	4	5
$\epsilon^A_{st}(1,q)$	0.04 (0.01)	0.02 (0.02)	0.05 (0.03)	0.14 (0.04)	0.24 (0.04)
$\epsilon^D_{st}(1,q)$	0.01 (0.01)	0.02 (0.03)	0.04 (0.05)	0.06 (0.07)	0.10 (0.10)
$\epsilon^A_{st}(2,q)$	0.03 (0.01)	0.02 (0.02)	0.04 (0.02)	0.08 (0.05)	0.14 (0.06)
$\epsilon^D_{st}(2,q)$	0.01 (0.01)	0.03 (0.03)	0.04 (0.05)	0.06 (0.07)	0.08 (0.09)

Table 2: The structure function error $\epsilon_{st}(j,q)$ of model A,D for the isotropic turbulent vorticity field as a function of q, for $j = 1, 2, 3$.

q	1	2	3	4	5
$\epsilon^A_{st}(1,q)$	0.06 (0.02)	0.04 (0.03)	0.16 (0.06)	0.38 (0.05)	0.60 (0.04)
$\epsilon^D_{st}(1,q)$	0.02 (0.01)	0.04 (0.03)	0.05 (0.05)	0.13 (0.08)	0.28 (0.09)

Table 3: The structure function error $\epsilon_{st}(j,q)$ of model A,D for the increment of MRW process as a function of q, for $j = 1$.

The evaluation methodology with moments is also able to detect model errors. It shows that long range correlations of wavelet harmonic coefficients are not well captured by foveal models. To circumvent this issue without increasing too much the model size, one can capture these long range correlations with another wavelet transform at each scale, which defines a scattering transform [30].

A Maximum Entropy Wavelet Covariance model

We explain how to compute the Lagrangian multiples of a maximum entropy stationary Gaussian random vector specified by wavelet covariance values, which is used for the model A.

To solve the maximum entropy problem, we follow [2] and derive a dual problem to minimize the entropy with respect to the Lagrangian multiples. To simplify the deviation, we assume that X is a zero-mean stationary process, and we consider only the translation group G. The wavelet transform is indexed by $v = (\lambda, u)$. The maximum-entropy wavelet covariance model can be written

$$\tilde{p}(x) = \frac{1}{Z} \exp\Big(-\frac{1}{2} \sum_{(v,v')\in E} \beta_{v,v'} \sum_{g\in G} x \star \psi_\lambda(u-g)\, x \star \psi^*_{\lambda'}(u'-g) \Big).$$

The sum over E is real-valued because $\beta_{v,v'} = \beta^*_{v',v}$.

With the Parseval formula, the sum over translations in G can be written in the Fourier domain as a sum over frequencies

$$\tilde{p}(x) = \frac{1}{Z} \exp\Big(-\frac{1}{2d} \sum_{(v,v')\in E} \beta_{v,v'} \sum_{\omega} |\widehat{x}(\omega)|^2 \widehat{\psi}_\lambda(\omega) \widehat{\psi}^*_{\lambda'}(\omega)\, e^{-i\omega.(u-u')} \Big).$$

The power spectrum $\widetilde{P}(\omega)$ of $\tilde{p}$ is

$$\widetilde{P}(\omega) = \frac{1}{d}\int |\hat{x}(\omega)|^2 \tilde{p}(x)dx = \Big(\sum_{(v,v')\in E} \beta_{v,v'}\, \widehat{\psi}_\lambda(\omega)\, \widehat{\psi}^*_{\lambda'}(\omega) e^{-i\omega.(u-u')}\Big)^{-1}.$$

To find the optimal Lagrange multipliers, we minimize the entropy of $\tilde{p}$ with respect to all the dual variables $\beta_{v,v'}$. The entropy has a closed form,

$$H(\tilde{p}) = \frac{1}{2}\sum_{(v,v')\in E} \beta_{v,v'} \sum_{\omega} \widetilde{P}(\omega)\widehat{\psi}_\lambda(\omega)\widehat{\psi}_{\lambda'}(\omega)^* e^{-i\omega.(u-u')} + \log(Z). \tag{43}$$

The partition function Z also has a closed form,

$$\log(Z) = \frac{1}{2}\sum_{\omega} \log \widetilde{P}(\omega) + \frac{d}{2}\log(2\pi).$$

Note that the $\beta_{v,v'}$ should be constrained so that $\widetilde{P}(\omega) > 0, \forall \omega$. This constrained optimization problem is addressed by setting the entropy loss to infinity if any condition is violated. The derivative of $H(\tilde{p})$ with respect to $\beta_{v,v'}$ is calculated from (43). We can thus use an unconstrained optimization solver L-BFGS to solve this problem from the real and imaginary parts of $\beta_{v,v'}$ and $\beta_{v',v}$. The optimal solution gives the power spectrum $\widetilde{P}(\omega)$ of the Gaussian model A. Samples of model A are obtained by sampling a stationary Gaussian random vector whose power spectrum is $\widetilde{P}(\omega)$.

B Proof of Theorem 2.1

We show that for the rectifier $\rho(a) = \max(a, 0)$,

$$\|\widehat{\mathcal{H}}(z) - \widehat{\mathcal{H}}(z')\|^2 \leq \frac{1}{4}\,|z - z'|^2, \quad \forall (z, z') \in \mathbb{C}^2. \tag{44}$$

By definition, $\mathcal{H}(z) = \{\rho(\mathrm{Real}(e^{i\alpha} z))\}_{\alpha\in[0,2\pi]}$. Since $\|\widehat{\mathcal{H}}(z) - \widehat{\mathcal{H}}(z')\|^2 = \|\mathcal{H}(z) - \mathcal{H}(z')\|^2$, we do the proof on $\|\mathcal{H}(z) - \mathcal{H}(z')\|^2$. We prove the theorem by first showing that for a rectifier $\rho(a) = \max(a, 0)$ and any $(a, a') \in \mathbb{R}^2$

$$|\rho(a) - \rho(a')| + |\rho(-a) - \rho(-a')| = |a - a'|.$$

Indeed $\rho(a) = a$ and $\rho(-a) = 0$ or $\rho(a) = 0$ and $\rho(-a) = -a$. The equality is verified by considering separately the cases where a and a' have same or different signs. For $a' = 0$ we get

$$|\rho(a)|^2 + |\rho(-a)|^2 = |a|^2 \tag{45}$$

and for any $a' \in \mathbb{R}$

$$\frac{1}{2}|a - a'|^2 \leq |\rho(a) - \rho(a')|^2 + |\rho(-a) - \rho(-a')|^2 \leq |a - a'|^2. \tag{46}$$

For any $z \in \mathbb{C}$, $\text{Real}(e^{i(\alpha+\pi)}z) = \text{Real}(-e^{i\alpha}z)$ and $\text{Real}(e^{i(\alpha+\pi/2)}z) = \text{Imag}(e^{i\alpha}z)$, so (45) and (46) imply that

$$\sum_{n=0}^{3} |\rho(\text{Real}(e^{i(\alpha+n\pi/2)}z))|^2 = |z|^2.$$

and

$$\frac{1}{2}|z-z'|^2 \leq \sum_{n=0}^{3} |\rho(\text{Real}(e^{i(\alpha+n\pi/2)}z)) - \rho(\text{Real}(e^{i(\alpha+n\pi/2)}z'))|^2 \leq |z-z'|^2.$$

Integrating both of these double inequalities over $\alpha \in [0, \pi/2]$ gives

$$\frac{1}{2\pi}\int_0^{2\pi} |\rho(\text{Real}(e^{i\alpha}z))|^2 d\alpha = \frac{1}{4}|z|^2,$$

and

$$\frac{1}{8}|z-z'|^2 \leq \frac{1}{2\pi}\int_0^{2\pi} |\rho(\text{Real}(e^{i\alpha}z)) - \rho(\text{Real}(e^{i\alpha}z'))|^2 d\alpha \leq \frac{1}{4}|z-z'|^2,$$

which proves (44).

C Proof of Theorem 2.2

Property (22) is proved similarly to (13) by translating X by $\tau \in \Lambda_d$ and by using its stationarity to verify that $\text{Cov}([\widehat{X}(\omega)]^k, [\widehat{X}(\omega')]^{k'}) = e^{i(k\omega - k'\omega').\tau}\,\text{Cov}([\widehat{X}(\omega)]^k, [\widehat{X}(\omega')]^{k'})$ so both terms vanish if $k\omega \neq k'\omega'$.

Property (25) is verified for $\omega \neq 0$ by decomposing nominator and denominator with $\text{Cov}(A,B) = \mathbb{E}(AB^*) - \mathbb{E}(A)\,\mathbb{E}(B^*)$. Property (23) uses the same decomposition and the fact that $\mathbb{E}(\widehat{X}(\omega)) = 0$ and

$$\mathbb{E}([\widehat{X}(\omega)]^k) = 0 \ \text{ if } \ k\omega \neq 0\ .$$

This is proved again by translating X by $\tau \in \Lambda_d$, which transforms $\widehat{X}(\omega)$ into $e^{-i\tau.\omega}\widehat{X}(\omega)$. Since X is stationary, it does not modify this expected value and hence $\mathbb{E}([\widehat{X}(\omega)]^k) = e^{-ik\tau.\omega}\mathbb{E}([\widehat{X}(\omega)]^k)$. Since this is valid for any $\tau \in \Lambda_d$, $\mathbb{E}([\widehat{X}(\omega)]^k) = 0$ if $k\omega \neq 0$ modulo 2π in dimension r.

Property (23) for $\omega = 0$ and property (24) are also proved by decomposing $\text{Cov}(A,B) = \mathbb{E}(AB^*) - \mathbb{E}(A)\,\mathbb{E}(B^*)$ and showing that

$$\mathbb{E}([\widehat{X}(0)]^{2k}) = \mathbb{E}(|\widehat{X}(0)|) \ \ , \ \ \mathbb{E}([\widehat{X}(0)]^{2k+1}) = \mathbb{E}(\widehat{X}(0)) = d\,\mathbb{E}(X(u))\ \ . \tag{47}$$

This results from the fact that $\widehat{X}(0) = \sum_{u\in\Lambda_d} X(u)$ is real so $[\widehat{X}(0)]^{2k} = |\widehat{X}(0)|$ and $[\widehat{X}(0)]^{2k+1} = \widehat{X}(0)$.

If X is Gaussian then $\widehat{X}(\omega)$ and $\widehat{X}(\omega')$ are independent if $\omega \neq \omega'$ because they are Gaussian random variables with a zero covariance. It results that $[\widehat{X}(\omega)]^k$ and $[\widehat{X}(\omega')]^{k'}$ are also independent for any $(k,k') \in \mathbb{Z}^2$ and hence have a zero covariance if $\omega \neq \omega'$. If $\omega = \omega'$ and $k \neq k'$ then (22) proves that their covariance remains zero. It results that $\mathrm{Cov}([\widehat{X}(\omega)]^k, [\widehat{X}(\omega')]^{k'}) \neq 0$ only if $(\omega, k) = (\omega', k')$ and hence that the covariance is diagonalized. If $\omega \neq 0$ then $\widehat{X}(\omega)$ is a zero-mean complex Gaussian random variable whose real and imaginary parts are not correlated. As a result $\mathbb{E}(|\widehat{X}(\omega)|)^2 / \mathbb{E}(|\widehat{X}(\omega)|^2) = \pi/4$.

D Bump Steerable wavelet

We review the bump steerable wavelet introduced in [4]. It is illustrated in Figure 1 and provides a sparse representation of images with oriented structures. A general way of constructing steerable wavelets is to use the polar coordinate in the Fourier frequency domain [16].

We specify the bump steerable wavelet along the radius $|\omega|$ and angle $arg(\omega)$ for $\omega = |\omega| e^{i \cdot arg(\omega)}$. We assume that the number of angles L is even and its central frequency is $\lambda = (\xi_0, 0)$. Its formula is

$$\widehat{\psi}(\omega) = c \cdot \exp\Big(\frac{-(|\omega| - \xi_0)^2}{\xi_0^2 - (|\omega| - \xi_0)^2}\Big)\, 1_{[0, 2\xi_0]}(|\omega|) \cdot \cos^{L/2-1}(arg(\omega)) 1_{arg(\omega) < \frac{\pi}{2}},$$

where c is a normalization constant. The radial function along $|\omega|$ is a bump function which is a compact-support approximation of a Gaussian window. The same angular window function along $arg(\omega)$ is used in [15].

As in [4], the father wavelet ϕ is an isotropic Gaussian window function,

$$\widehat{\phi}(\omega) = \exp\Big(- \frac{|\omega|^2}{2\sigma^2} \Big).$$

We choose $\xi_0 = 1.7\pi$, $\sigma = 0.248 \times 2^{-0.55} \xi_0$ and $c = 1.29^{-1} 2^{\frac{L}{2}-1} \frac{(\frac{L}{2}-1)!}{\sqrt{(\frac{L}{2})(L-2)!}}$. These hyperparameters are also used in [4]. For the wavelet transform $\mathcal{W}x$ with oversampling, we obtain numerically the frame constants $A_{\mathcal{W}} = 2.0$ and $B_{\mathcal{W}} = 4.6$ for $d = 128^2$, $J = 5$ and $L = 16$. It is therefore a complete and stable representation. These mother and father wavelets also satisfy $\psi(u_1, -u_2) = \psi(u_1, u_2)$ and $\phi(u_1, -u_2) = \phi(u_1, u_2)$.

E Proof of Theorem 4.1

Observe first that μ_0 is invariant to the action of any $g \in G$. Indeed, the probability measure μ_0 of a Gaussian white noise is uniform on an ball of $\mathbb{R}^d$ centered in 0. As a result μ_0 is invariant to the action of any linear unitary operator and hence invariant to the action of any $g \in G$ which is linear and unitary. Since μ_0 is invariant to the action of g, if the gradient descent is covariant to the action of g then the derivations of Theorem 3.4 in [11] prove that the probability measure μ_t of x_t is invariant to g for $t \geq 0$.

We thus need to show that the L-BFGS gradient-descent algorithm is covariant to the action $g \in G$ at each iteration (before stopped by line-search conditions). Let $\tilde{x}_t$

and x_t be two sequences generated by the gradient descent from the initial conditions $\tilde{x}_0$ and x_0. The covariance to the action of g means that if $\tilde{x}_0 = g.x_0$ then $\tilde{x}_t = g.x_t$ for any $t \geq 0$. Although invariant properties of a gradient-based algorithm are well studied in the optimization literature [25, 31], we give below a prove for completeness.

L-BFGS is a variant of the quasi-Newton method BFGS. It estimates the inverse Hessian matrix of the objective function f based on

$$s_t = x_{t+1} - x_t \text{ and } y_t = \nabla f(x_{t+1}) - \nabla f(x_t).$$

Assume there is a limited memory size m to compute an approximation H_t^m of the Hessian, from s_{t-a} and y_{t-a} for $1 \leq a \leq m$. At each iteration t, the algorithm chooses an appropriate step-size α_t satisfying strong Wolfe conditions [25] and updates x_t with

$$x_{t+1} = x_t - \alpha_t H_t^m \nabla f(x_t).$$

Since $f(x) = \|\widetilde{K}_{\mathcal{R}x} - \widetilde{K}_{\mathcal{R}\bar{x}}\|_{E_G}^2$ and $\widetilde{K}_{\widehat{H}\mathcal{W}g.x} = \widetilde{K}_{\widehat{H}\mathcal{W}x}$, it results that $f(g.x) = f(x)$. It implies that

$$\nabla f(g.x) = g.\nabla f(x). \tag{48}$$

We prove the covariance property by induction. We suppose that $\tilde{x}_{t'} = g.x_{t'}$ for $t' \leq t$ in order to prove that $\tilde{x}_{t+1} = g.x_{t+1}$. By definition,

$$\tilde{x}_{t+1} = \tilde{x}_t - \tilde{\alpha}_t \tilde{H}_t^m \nabla f(\tilde{x}_t).$$

The induction assumption implies that $\tilde{s}_{t'} = g.s_{t'}$ and $\tilde{y}_{t'} = g.y_{t'}$ for $t' \leq t$. It is therefore sufficient to verify that

$$\tilde{H}_t^m = g.H_t^m.g^T, \quad \tilde{\alpha}_t = \alpha_t, \quad \forall t \geq 0$$

Each H_t^m is defined recursively by

$$H_t^a = V_{t+a-m-1}^T H_t^{a-1} V_{t+a-m-1} + \rho_{t+a-m-1} s_{t+a-m-1} s_{t+a-m-1}^T,$$

for $1 \leq a \leq m$ and for all $t \geq 0$

$$V_t = Id - \rho_t\, s_t\, y_t^T \text{ and } \rho_t = (s_t^T\, y_t)^{-1}.$$

We thus have $\tilde{V}_t = g.V_t.g^T$. By convention, $H_0^0 = Id$ and $H_t^0 = (s_{t-1}^T\, y_{t-1})^{-1} Id$ for $t \geq 1$. We thus conclude that $\tilde{H}_t^m = gH_t^m g^T$. Note that if $t \leq m$, the index $t+a-m-1$ is negative, and in this case $V_{t+a-m-1} = Id$ and $\rho_{t+a-m-1} = 0$.

To verify that $\tilde{\alpha}_t = \alpha_t$, we observe that the strong Wolfe-line search conditions are invariant to any linear unitary $g \in G$. Indeed these conditions are invariant to any affine operator [25].

F Proof of Proposition 5.1

If p is invariant to an action of G then $M_{\widehat{H}\mathcal{W}}$ and $K_{\widehat{H}\mathcal{W}}$ are not modified if X is transformed into $g.X$. Similarly, we verify from (3) and (4) that $\widetilde{M}_{\widehat{H}\mathcal{W}}$ and $\widetilde{K}_{\widehat{H}\mathcal{W}}$ are not modified if $\bar{x}$ is transformed into $g.\bar{x}$. The proposition results are obtained by relating $[g.x \star \psi_\lambda]^k$ with $[x \star \psi_\lambda(u)]^k$ and applying this relation to $x = X$ and $x = \bar{x}$.

(i). If $g.x = -x$ then

$$[g.x \star \psi_\lambda(u)]^k = (-1)^k [x \star \psi_\lambda(u)]^k. \tag{49}$$

For $v = (\lambda, k, u)$ we thus derive that $M_{\widehat{H}\mathcal{W}}(v) = (-1)^k M_{\widehat{H}\mathcal{W}}(v)$ and $\widetilde{M}_{\widehat{H}\mathcal{W}}(v) = (-1)^k \widetilde{M}_{\widehat{H}\mathcal{W}}(v)$ and thus vanish if k is odd. Similarly, for $v' = (\lambda', k', u')$ we verify that $K_{\widehat{H}\mathcal{W}}(v, v') = (-1)^{k+k'} K_{\widehat{H}\mathcal{W}}(v, v')$ and $\widetilde{K}_{\widehat{H}\mathcal{W}\bar{x}}(v, v') = (-1)^{k+k'} \widetilde{K}_{\widehat{H}\mathcal{W}\bar{x}}(v, v')$ which vanishes if $k + k'$ is odd.

(ii) If $g.x(u) = x(-u)$ then

$$[g.x \star \psi_\lambda(u)]^k = [x \star \psi_\lambda(u)]^{k*} \tag{50}$$

because $\psi_\lambda(-u) = \psi_\lambda^*(u)$ and x is real. Since $\widehat{h}(k)$ is real, $M_{\widehat{H}\mathcal{W}}(v) = \widehat{h}(k)\,\mathbb{E}([X \star \psi_\lambda(u)]^k)^* = M_{\widehat{H}\mathcal{W}}(v)^*$ is real. The same property applies to $\widetilde{M}_{\widehat{H}\mathcal{W}}(v)$. Applying (50) to $K_{\widehat{H}\mathcal{W}}(v, v')$ and $\widetilde{K}_{\widehat{H}\mathcal{W}\bar{x}}(v, v')$ also proves that they are real.

G Proof of Theorem 5.1

Since $\lambda = 2^{-j} r_{-\ell}\xi$ the index of $\mathcal{R} = \widehat{\mathcal{H}\mathcal{W}}$ can be written $v = (j, \ell, k, u)$. We compute the covariance $K_{\widehat{\mathcal{H}}\mathcal{W}}(v, v')$ for $v' = (j', \ell', k', u')$ with $u' = u$. Since the distribution is stationary the covariance value remains the same if we set $u = u' = 0$.

To prove the theorem we first observe that if $g.x = r_\eta x$ then

$$[g.x \star \psi_\lambda(0)]^k = [x \star \psi_{r_\eta \lambda}(0)]^k. \tag{51}$$

If X is isotropic then $K_{\widehat{H}\mathcal{W}}(v, v')$ and $\widetilde{K}_{\widehat{H}\mathcal{W}\bar{x}}(v, v')$ are invariant by rotations. The covariance property (51) thus implies that they are a function of the difference of angles $\ell - \ell'$. It is therefore a periodic convolution kernel along angles. It results that the angular dependence is diagonalized by a Fourier transform along the angle variable, which proves (39) for $\mathcal{R} = \mathcal{F}_\ell\, \widehat{\mathcal{H}\mathcal{W}}$.

A central reflection is a rotation by π. Since the distribution of X is invariant to all rotations it is invariant to a central reflection. Property (ii) of Proposition (5.1) implies that $K_{\widehat{H}\mathcal{W}}$ is real.

Let us consider the line reflection g relatively to the horizontal axis. Since $\psi(u_1, -u_2) = \psi(u_1, u_2)$ and $\phi(u_1, -u_2) = \phi(u_1, u_2)$ we verify that for any $\lambda = 2^{-j} r_{-\ell}\xi$ and $\lambda' = 2^{-j} r_\ell \xi$ we have

$$\psi_\lambda(u_1, u_2) = \psi_{\lambda'}(u_1, -u_2). \tag{52}$$

As a result, if $\lambda = 2^{-j} r_{-\ell}\xi$ then

$$[(g.x) \star \psi_\lambda(0)]^k = [x \star \psi_{\lambda'}(0)]^k \quad \text{with} \quad \lambda' = 2^{-j} r_\ell \xi. \tag{53}$$

If $K_{\widehat{\mathcal{H}}\mathcal{W}}$ is invariant to line reflections then (53) implies that $K_{\widehat{\mathcal{H}}\mathcal{W}}$ is not modified when (ℓ, ℓ') is transformed into $(-\ell, -\ell')$. Since $K_{\widehat{\mathcal{H}}\mathcal{W}}$ is real and invariant to a change of sign of (ℓ, ℓ') it results that the Fourier coefficients of $K_{\mathcal{F}_\ell \widehat{\mathcal{H}}\mathcal{W}}$ are also real.

References

[1] Jaynes, E. T. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[2] Cover, T. M and Thomas, J. A. *Elements of Information Theory.* Wiley-Interscience, New York, NY, USA, 2006.

[3] Priestley, M. B. *Spectral analysis and time series.* Academic Press, London ; New York :, 1981.

[4] Mallat, S, Zhang, S, and Rochette, G. Phase harmonic correlations and convolutional neural networks. *Information and Inference: A Journal of the IMA*, 11 2019.

[5] Wendt, H, Abry, P, and Jaffard, S. Bootstrap for empirical multifractal analysis. *IEEE Signal Processing Magazine*, 24(4):38–48, July 2007.

[6] Farge, M, Schneider, K, and Kevlahan, N. Non-gaussianity and coherent vortex simulation for two-dimensional turbulence using an adaptive orthogonal wavelet basis. *Physics of Fluids*, 11(8):2187–2201, 1999.

[7] Grossmann, A, Kronland-Martinet, R, and Morlet, J. Reading and Understanding Continuous Wavelet Transforms. In Combes, J.-M, Grossmann, A, and Tchamitchian, P, editors, *Wavelets. Time-Frequency Methods and Phase Space*, page 2, 1989.

[8] Portilla, J and Simoncelli, E. P. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, Oct 2000.

[9] Ustyuzhaninov, I, Brendel, W, Gatys, L, and Bethge, M. What does it take to generate natural textures? In *International Conference on Learning Representations*, Apr 2017.

[10] Lustig, R. Microcanonical monte carlo simulation of thermodynamic properties. *The Journal of Chemical Physics*, 109(20):8816–8828, 1998.

[11] Bruna, J and Mallat, S. Multiscale sparse microcanonical models. *Mathematical Statistics and Learning*, 1:257–315, 2018.

[12] Zhu, S. C, Wu, Y. N, and Mumford, D. Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9(8):1627–1660, 1997.

[13] Cramér, H. *Mathematical Methods of Statistics (PMS-9).* Princeton University Press, 1999.

[14] Rao, T. S and Gabr, M. *An introduction to bispectral analysis and bilinear time series models*, volume 24. Springer Science & Business Media, 2012.

[15] Simoncelli, E. P and Freeman, W. T. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *Proceedings., International Conference on Image Processing*, volume 3, pages 444–447 vol.3, Oct 1995.

[16] Unser, M and Chenouard, N. A unifying parametric framework for 2d steerable wavelet transforms. *SIAM Journal on Imaging Sciences*, 6(1):102–135, 2013.

[17] Mallat, S. *A Wavelet Tour of Signal Processing: The Sparse Way, 3rd Edition.* Academic Press, 2001.

[18] Schneider, K, Ziuber, J, Farge, M, and Azzalini, A. Coherent vortex extraction and simulation of 2d isotropic turbulence. *Journal of Turbulence*, (7):N44, 2006.

[19] Andreux, M. *Foveal Autoregressive Neural Time-Series Modeling.* PhD thesis, École normale supérieure, 2018.

[20] Bortoli, V. D, Desolneux, A, Galerne, B, and Leclaire, A. Macrocanonical models for texture synthesis. *Proc. of Int. Conf. on scale Space and Variational Methods in Computer vision*, June 2019.

[21] Dembo, A and Zeitouni, O. *Large Deviations Techniques and Applications.* Johns and Bartett Publishers, Boston, 1993.

[22] Deuschel, J, Stroock, D, and Zession, H. Microcanonical distributions for lattice gases. *Commun. Math. Phys.*, 139:83–101, 1991.

[23] Georgii, H.-O. *Gibbs measures and phase transitions*, volume 9. Walter de Gruyter, 2011.

[24] Lustig, R. Microcanonical monte carlo simulation of thermodynamic properties. *The Journal of Chemical Physics*, 109(20):8816–8828, 1998.

[25] Nocedal, J and Wright, S. J. *Numerical Optimization.* Springer, New York, NY, USA, second edition, 2006.

[26] Mallat, S. Foveal detection and approximation for singularities. *Applied and Computational Harmonic Analysis*, 14(2):133–180, 2003.

[27] Bacry, E, Delour, J, and Muzy, J.-F. Multifractal random walk. *Physical Review E*, 64(2):026103, 2001.

[28] Van'Yan, P. Structure function of the velocity field in turbulent flows. *Journal of Experimental and Theoretical Physics*, 82(3):580–586, 1996.

[29] Farge, M, Kevlahan, N, Perrier, V, and Schneider, K. Turbulence analysis, modelling and computing using wavelets. *Wavelets in Physics*, pages 117–200, 1999.

[30] Leonarduzzi, R, Rochette, G, and Mallat, S. Time series models with wavelet scattering harmonics. *in preparation*, 2019.

[31] Fletcher, R. *Practical Methods of Optimization, 2nd Edition.* Wiley-Interscience, New York, NY, USA, 1987.