Math. Stat. Learn. 1 (2018), 257–315
DOI 10.4171/MSL/7

**Mathematical Statistics and Learning**


# Multiscale sparse microcanonical models

Joan Bruna and Stéphane Mallat

**Abstract.** We study approximations of non-Gaussian stationary processes having long range correlations with microcanonical models. These models are conditioned by the empirical value of an energy vector, evaluated on a single realization. Asymptotic properties of maximum entropy microcanonical and macrocanonical processes and their convergence to Gibbs measures are reviewed. We show that the Jacobian of the energy vector controls the entropy rate of microcanonical processes.

Sampling maximum entropy processes through MCMC algorithms require too many operations when the number of constraints is large. We define microcanonical gradient descent processes by transporting a maximum entropy measure with a gradient descent algorithm which enforces the energy conditions. Convergence and symmetries are analyzed. Approximations of non-Gaussian processes with long range interactions are defined with multiscale energy vectors computed with wavelet and scattering transforms. Sparsity properties are captured with $\mathbf{l}^1$ norms. Approximations of Gaussian, Ising and point processes are studied, as well as image and audio texture synthesis.

*Mathematics Subject Classification* (2010). 49Q20, 53A07, 62G07, 62M45, 82B20; 82B28, 82B80.

*Keywords.* Macrocanonical, microcanonical, wavelet, scattering, texture.

## 1. Introduction

Building probabilistic models of large systems of interacting variables that can be efficiently estimated from data is a core problem in statistical physics, machine learning and signal processing. We consider the estimation of the probability measure of stationary processes $X(u)$ on the infinite grid $u \in \mathbb{Z}^\ell$ given a single realization $\bar{x}(u)$, observed over a finite domain $u \in \Lambda_d \subset \mathbb{Z}^\ell$ of cardinality $d$. For $\ell = 2$ and $\ell = 1$, such processes provide models of image and audio textures. Given a piece of texture over $\Lambda_d$, we may want to synthesize similar texture examples by sampling the resulting probability model. Building probability models from a single observation is also needed in finance and in many physical problems, such as geophysics exploration or fluid dynamics. These estimations rely on the ability to build low-dimensional approximations of the underlying stationary measure. This paper introduces microcanonical sparse multiscale models, which can take into account non-Gaussian phenomena and long range interactions.

In his seminal paper, Jaynes [27] interprets statistical physics as an inference of a probability distribution from partial measurements, by maximizing its entropy. In Jaynes words [27], maximizing the entropy of a probability distribution "is maximally noncommittal with regard to missing information." Macrocanonical models are maximum entropy distributions conditioned on the expected value of a vector of potential energies. They are used in large classes of stochastic models [24] and will thus be our departure point.

Since we only know a single realization $\bar{x}(u)$ of $X(u)$ in $\Lambda_d$, the expected value of stationary energies are estimated by the average potential energy vector $\Phi_d(\bar{x})$ of $\bar{x}$ in the domain $\Lambda_d$ of size $d$. When $d$ is sufficiently large, weak ergodicity assumptions imply that $\Phi_d(X)$ concentrates near the empirical energy vector $\Phi_d(\bar{x})$ with high probability. A microcanonical model is a probability measure supported over the microcanonical set of all $x$ having nearly the same energy: $\|\Phi_d(x) - \Phi_d(\bar{x})\| \leq \epsilon$. Maximum entropy microcanonical models have a uniform density over this set. Under appropriate hypotheses, the Boltzmann equivalence principle states that a maximum entropy microcanonical model converges to the same Gibbs measure as the macrocanonical model, when $d$ goes to $\infty$. Section 2.4 reviews these results.

Microcanonical models exist with mild assumptions, even-though macrocanonical distributions may not exist, particularly for signals $x$ having strong sparsity properties. We thus consider these models not as approximations of macrocanonical models, which may not exist, but as stochastic models in their own sake. Section 3 relates their entropy rate to their energy vector. Sampling micro and macrocanonical measures is a classic problem in statistical mechanics, typically approached with MCMC algorithms or Langevin Dynamics [6, 15] or variational methods [45]. Their numerical effectiveness on high-dimensional problems is hindered by the slow mixing speed of the Markov Chain [15], which limits their applications. To avoid this computational issue, we introduce an alternative class of microcanonical models where the Markov chain is replaced by a gradient flow resulting from the microcanonical energy vector. A microcanonical gradient descent model begins from a high entropy measure and computes a progressive transport of this measure with gradient steps, towards the microcanonical set. Similar algorithms have been applied to texture synthesis [23] with deep convolutional neural networks. Section 3 studies their convergence to a microcanonical set. Although the gradient descent transport does not converge to a maximum entropy measure, we prove that it preserves an important subset of symmetries which is specified.

A major issue is to specify energy vectors $\Phi_d$ providing accurate microcanonical gradient descent approximations of non-Gaussian processes with long range interactions. Section 4 introduces energy vectors which take into account long range interactions by separating scales with wavelet transforms. Non-Gaussian properties are captured with $\mathbf{l}^1$ norms which measure the sparsity of wavelet coefficients.

These energy vectors are augmented with wavelet scattering coefficients, providing information on the geometry of sparse wavelet coefficients [10, 33].

Section 5 studies the approximation of Gaussian, Ising and point processes, with microcanonical gradient descent models computed with wavelet and scattering energy vectors. For Ising, the wavelet scale separation is closely related to the Wilson renormalization group approach [5]. We show that scattering microcanonical model can also give good perceptual approximations of large classes of image and audio textures.

**Notation.** We use cursive captial letters $\mathcal{A}, \mathcal{B}, \ldots$ to denote sets, small capital letters $x, y, \ldots$ to denote vectors, capitals $X, Y, Z$ to denote random processes, and capital letters $E, H, \Phi, \ldots$ to denote operators and functions. $\widehat{x}$ denotes the Fourier transform of $x$. $\|x\|$ denotes the Euclidean norm of $x$.

## 2. Microcanonical and macrocanonical models

We consider a stationary process $X(u)$ taking its values in an interval $\mathcal{J} \subseteq \mathbb{R}$ for all $u \in \mathbb{Z}^\ell$. We denote by $\mu$ the probability measure of this stationary process. We write $\mathbb{E}_\mu(f(x))$ the expected value of $f(X)$ or $\mathbb{E}_p(f(x))$ if $\mu$ has a density $p$. Let $\Lambda_d \subset \mathbb{Z}^\ell$ be a cube with $d$ grid points and $\mathcal{J}_d^\Lambda$ the product domain. Let $\bar{x} \in \mathcal{J}_d^\Lambda$ be a realization of $X$ restricted to $\Lambda_d$. Microcanonical models described in Section 2.1 are probability densities conditioned on a $K$-dimensional energy vector $\Phi_d(\bar{x})$. Section 2.2 reviews the properties of macrocanonical models which have a maximum entropy conditioned on $\mathbb{E}_\mu(\Phi_d(x))$. We concentrate on shift-invariant energies $\Phi_d$ introduced in Section 2.3, to define stationary maximum entropy processes. Section 2.4 reviews the resulting convergence properties of micro and macrocanonical models towards the same Gibbs measures. In statistical physics terms, it amounts to verify the Boltzmann equivalence principle in the thermodynamical limit, for lattice gaz models. We shall then see that microcanonical models are also interesting in their own sake, even in regimes where macrocanonical models do not exist.

### 2.1. Maximum entropy microcanonical models.

A microcanonical model is computed from $y = \Phi_d(\bar{x})$. To estimate the measure $\mu$ of a stationary $X$ from a single realization, we need ergodicity assumptions. We assume that $\Phi_d(X)$ concentrates with high probability around $\mathbb{E}_\mu(\Phi_d(x))$ when $d$ goes to $\infty$:

$$\forall \epsilon > 0\,, \quad \lim_{d\to\infty} \mathrm{Prob}_\mu\big(\|\Phi_d(X) - \mathbb{E}_\mu(\Phi_d(x))\| \leq \epsilon\big) = 1\,. \tag{1}$$

If there exists $C > 0$ such that $\|\mathbb{E}_\mu(\Phi_d(x))\| \leq C$ then this convergence in probability is implied by a mean-square convergence:

$$\lim_{d\to\infty} \mathbb{E}_\mu\big(\|\Phi_d(x) - \mathbb{E}_\mu(\Phi_d(x))\|^2\big) = 0\,. \tag{2}$$

The microcanonical set of width $\epsilon$ associated to $y = \Phi_d(\bar{x})$ is

$$\Omega_{d,\epsilon} = \big\{x \in \mathcal{I}_d^\Lambda : \|\Phi_d(x) - y\| \leq \epsilon\big\}\,.$$

The concentration property (1) implies that when $d$ goes to $\infty$, $X$ belongs to microcanonical sets $\Omega_{d,\epsilon}$ of width $\epsilon = \epsilon(d)$ converging to 0, with a probability converging to 1. In other words, (1) guarantees that the support of the measure $\mu$ is mostly concentrated in $\Omega_{d,\epsilon}$ for large $d$.

The differential entropy of a probability distribution $\mu$ which admits a density $p(x)$ relatively to the Lebesgue measure is

$$H(\mu) := -\int p(x)\,\log p(x)\,dx\,. \tag{3}$$

A maximum entropy microcanonical model $\mu^{\mathrm{mi}}(d,\epsilon,y)$ was defined by Boltzmann as the maximum entropy distribution supported in $\Omega_{d,\epsilon}$. We usually define $\Phi_d(x)$ so that $\Omega_{d,\epsilon}$ is compact. It results the maximum entropy distribution has a uniform density $p_{d,\epsilon}$:

$$p_{d,\epsilon}(x) := \frac{1_{\Omega_{d,\epsilon}}(x)}{\int_{\Omega_{d,\epsilon}} dx}\,. \tag{4}$$

Its entropy is therefore the logarithm of the volume of $\Omega_{d,\epsilon}$:

$$H(p_{d,\epsilon}) = -\int p_{d,\epsilon}(x)\,\log p_{d,\epsilon}(x)\,dx = \log\Big(\int_{\Omega_{d,\epsilon}} dx\Big)\,. \tag{5}$$

We thus face a fundamental trade-off when constructing microcanonical models. On the one hand, we seek representations $\Phi_d$ that satisfy a concentration property (1) to ensure that typical samples from $\mu$ are included in $\Omega_{d,\epsilon}$ with high probability, and hence typical for the microcanonical measure $\mu^{\mathrm{mi}}$. On the other hand, the sets $\Omega_{d,\epsilon}$ must not be too large to avoid having elements of $\Omega_{d,\epsilon}$ and hence typical samples of $\mu^{\mathrm{mi}}$ which are not typical for $\mu$. To obtain an accurate microcanonical model, the energy $\Phi_d$ must define microcanonical sets of minimum volume, while satisfying the concentration (1).

**2.2. Macrocanonical models.** Since $\Phi_d(X)$ concentrates close to $\mathbb{E}_\mu(\Phi_d(x))$ and $\bar{x}$ is a realization of $X$, one could expect that the maximum entropy distribution conditioned on $\Phi_d(\bar{x})$ converges to the maximum entropy distribution conditioned on $\mathbb{E}_\mu(\Phi_d(x))$ when $d$ goes to $\infty$. Section 2.3 studies conditions under which this

Boltzmann equivalence principle is verified. We begin by reviewing the properties of macrocanonical maximum entropy models conditioned on $\mathbb{E}_\mu(\Phi_d(x)) = y$. Let $\mathcal{M}(\mathcal{I}_d^\Lambda)$ denote the space of measures of $\mathcal{I}_d^\Lambda$.

A macrocanonical measure $\mu^{\mathrm{ma}}$ with density $p_{\mathrm{ma}}$ has a maximum entropy conditioned on $\mathbb{E}_{p_{\mathrm{ma}}}(\Phi_d(x)) = y$:

$$p_{\mathrm{ma}} \in \arg\max_{p \in \mathcal{A}_y} H(p)\,,$$

with

$$\mathcal{A}_y = \left\{ p \in \mathcal{M}(\mathcal{I}_d^\Lambda);\, \int_{\mathcal{I}_d^\Lambda} \Phi_d(x)\, p(x)\, dx = y \right\}. \tag{6}$$

The entropy is a concave function of $p$ whereas $\mathbb{E}_p(\Phi_d(x)) = y$ is a set of linear conditions over $p$. If $\Phi_d(x)$ is bounded over $\Omega_{d,\epsilon}$ then the set of densities $p$ which satisfy the moment conditions is compact. As a consequence, there exists a unique macrocanonical density $p_{\mathrm{ma}}$ which maximizes $H(p)$. It is obtained by minimizing the following Lagrangian

$$\mathcal{L}_d(p, \beta) = -H(p) + \langle \beta, \mathbb{E}_p(\Phi_d(x)) - y \rangle\,, \tag{7}$$

also called free energy in statistical physics. The Lagrange multipliers $\beta = \{\beta_k\}_{k \le K}$ are adjusted so that the moment condition (6) is satisfied. The density which minimizes (7) can be written as an exponential family

$$p_{\mathrm{ma}}(x) = \mathcal{Z}^{-1} \exp\big( - \langle \beta, \Phi_d(x) \rangle \big)\,, \tag{8}$$

where $\mathcal{Z}$ guarantees that $\int p_{\mathrm{ma}}(x)\, dx = 1$ and hence

$$\mathcal{Z} = \int_{\mathcal{I}_d^\Lambda} \exp\big(\langle \beta, \Phi_d(x) \rangle\big)\, dx\,. \tag{9}$$

A direct calculation shows that the resulting maximum entropy is

$$H(p_{\mathrm{ma}}) = -\log \mathcal{Z} + \langle \beta, y \rangle\,. \tag{10}$$

If the probability measure of the restriction of $X$ to $\Lambda_d$ has a density $p$ relatively to the Lebesgue measure, then we can also verify that the Kullback–Liebler divergence

$$KL(p \| p_{\mathrm{ma}}) = \int_{\Lambda_d} p(x)\, \log \frac{p_{\mathrm{ma}}(x)}{p(x)}\, dx$$

satisfies

$$KL(p \| p_{\mathrm{ma}}) = H(p_{\mathrm{ma}}) - H(p) \ge 0\,. \tag{11}$$

Optimizing the interaction energy $\Phi_d$ thus amounts to minimizing the resulting maximum entropy $H(p_{\mathrm{ma}})$ [49] so that $H(p_{\mathrm{ma}}) = H(p)$ and hence $\mu^{\mathrm{ma}} = \mu$.

Note that it is not necessary to impose that $\Phi_d$ is bounded on $\mathcal{I}_d^\Lambda$. If there exists $\beta \in \mathbb{R}^K$ such that the distribution (8) satisfies the moment condition (6), then one can verify from (11) that $\mu^{\mathrm{ma}}$ is the unique maximum entropy distribution. However, if $\Phi_d$ is not bounded on $\mathcal{I}_d^\Lambda$ then there may not exist such a $\beta \in \mathbb{R}^K$. Indeed, the maximization of entropy defines a limit distribution over distributions which satisfy the moment constraints, but this limit may not satisfy the moment constraints anymore. One can construct such examples with high order moment conditions [44]. In this case the macrocanonical model does not exist although we may still define a microcanonical model.

**Macrocanonical estimation.** Given an energy vector $\Phi_d$, and desired moment constraints $y = \mathbb{E}_\mu[\Phi(x)]$, fitting macrocanonical models requires estimating $\mathbb{E}_{\mu^{\mathrm{ma}}}[\Phi_d(x)]$. This expectation can be estimated with MCMC algorithms such as Metropolis–Hastings, which sample the Gibbs distribution (8) to estimate $\mathbb{E}_{\mu^{\mathrm{ma}}}(\Phi_d(x))$ and iteratively update the Lagrange multipliers $\beta$ until $\mathbb{E}_{\mu^{\mathrm{ma}}}(\Phi_d(x))$ converges to $y$. However, when $d$ is large, this is numerically unfeasible because sampling a high-dimensional probability distribution is computationally dominated by the mixing time of the Markov Chain, which in generally has an exponential dependence on the data dimensionality [32].

**2.3. Shift equivariant and finite range potentials.** Microcanonical densities in (4) and macrocanonical densities in (8) depend on $\Phi_d$. These densities remain constant under any transformation of $x$ which leaves $\Phi_d(x)$ constant. Stationary densities are obtained with a $\Phi_d$ which is invariant to translations. It is calculated by averaging a potential vector which is equivariant to translations. We review simple examples with $\mathbf{l}^1$ and $\mathbf{l}^2$ norms. It illustrates convergence issues of micro and macrocanonical densities when $d$ goes to $\infty$, with sparse regimes where microcanonical models exist without macrocanonical models.

**Equivariant potentials.** For any $x \in \mathcal{I}^{\mathbb{Z}^\ell}$ we define a potential $Ux(u) \in \mathbb{R}^K$ for each $u \in \mathbb{Z}^\ell$. We write $T_\tau x(u) = x(u - \tau)$ a translation of $x$ by $\tau \in \mathbb{Z}^\ell$. A potential $U$ is shift-equivariant if

$$\forall (x, \tau) \in \mathcal{I}^{\mathbb{Z}^\ell} \times \mathbb{Z}^\ell \,,\; U T_\tau x = T_\tau U x \,.$$

The energy $\Phi_d(x)$ is computed from the restriction of $x$ in a square $\Lambda_d = [a, b]^\ell$. We extend $x$ over $\mathbb{Z}^d$ into a signal which is $b - a = d^{1/\ell}$ periodic along each of the $\ell$ generators of the grid $\mathbb{Z}^\ell$. With an abuse of notation we write $Ux$ the potential $U$ applied to the periodic extension of $x$ and

$$\Phi_d(x) = d^{-1} \sum_{u \in \Lambda_d} Ux(u) \,. \tag{12}$$

Observe that $\Phi_d(x) \in \mathbb{R}^K$ is invariant to periodic translations of $x$ in $\Lambda_d$ modulo $d^{1/\ell}$.

We say that $Ux$ has a finite range $\Delta$ if $Ux(u)$ only depends upon the values of $x(u')$ for $u - u' \in [-\Delta, \Delta]^\ell$. The resulting macrocanonical density (8) is a Markov Random Field over cliques $[u - \Delta, u + \Delta]^\ell$ around each $u$

$$p_{\rm ma}(x) = \mathcal{Z}^{-1} \exp\Big( - d^{-1} \sum_{u \in \Lambda_d} \langle \beta, Ux(u) \rangle \Big) . \tag{13}$$

To approximate random processes, we must choose $\Delta$ to be the integral scale beyond which structures become independent. When there are long range interactions as in turbulent flows, this integral scale may be very large. Before reviewing the general convergence properties of the resulting micro and macrocanonical densities we consider two important examples obtained with $\mathbf{l}^r$ norms.

**Convergence of $\mathbf{l}^r$ macro and microcanonical densities.** The potential $Ux(u) = |x(u)|^r$ for $u \in \mathbb{Z}$ defines an $\mathbf{l}^r$ norm energy over intervals $\Lambda_d = [1, d] \subset \mathbb{Z}$:

$$\Phi_d(x) = d^{-1}\, \|x\|_r^r = d^{-1} \sum_{u \in \Lambda_d} |x(u)|^r . \tag{14}$$

The macrocanonical measure with density $p_{\rm ma}$ defined by $\mathbb{E}_{p_{\rm ma}}(\Phi_d(x)) = y \geq 0$ is

$$p_{\rm ma}(x) = \mathcal{Z}^{-1} e^{-\beta\, d^{-1}\, \|x\|_r^r}$$

for some $\beta > 0$. It is the density of a vector of $d$ i.i.d random variables $X_d(u)$ having an exponential distribution $\propto e^{-\beta |z|^r}$.

A microcanonical density $p_{d,\epsilon,y}$ is uniform over $\Omega_{d,\epsilon} = \{x \in \mathbb{R}^d : |d^{-1} \|x\|_r^r - y| \leq \epsilon\}$, which is a thin shell around an $\mathbf{l}^r$ ball in $\mathbb{R}^d$. It is the density of a random vector $X_{d,\epsilon}$ defined on $\Lambda_d$. For a fixed $m > 0$, when $d$ goes to $\infty$ and $\epsilon$ goes to zero then the joint density of $X_{d,\epsilon}(1), \ldots, X_{d,\epsilon}(m)$ converges in total variation distance to i.i.d random variables having an exponential distribution $\propto e^{-\beta |z|^r}$ [4], and $\mathbb{E}(|X_{d,\epsilon}(u)|^r)$ converges to $y$. The microcanonical distribution thus converges to the macrocanonical distribution. This family of results has a long history, first proved in 1906 by Borel [7] for $r = 2$ and in 1987 by Diaconis and Freeman for $r = 1$ [18].

**Intersections of $\mathbf{l}^1$ and $\mathbf{l}^2$ balls.** The situation becomes more complex for the two-dimensional potential $Ux(u) = (|x(u)|^1, |x(u)|^2)$ which defines an energy $\Phi_d(x) = (d^{-1} \|x\|_1, d^{-1} \|x\|_2^2)$ over intervals $\Lambda_d = [1, d] \subset \mathbb{Z}$. We shall see that microcanonical models may exist without macrocanonical models.

One can verify that there exists a unique maximum entropy density $p_{\rm ma}$ conditioned on $\mathbb{E}_{p_{\rm ma}}(\Phi_d(x)) = y$ if and only if

$$1 \leq \frac{y_2}{y_1^2} \leq 2 ,$$

in which case there exists $\beta_1$ and $\beta_2$ such that

$$p_{\mathrm{ma}}(x) = \mathcal{Z}^{-1} e^{-d^{-1}\,(\beta_1 \|x\|_1 + \beta_2 \|x\|_2^2)} .$$

The microcanonical set $\Omega_{d,\epsilon} = \{x : \|\Phi_d(x) - y\| \leq \epsilon\}$ is a thin shell around the intersection of the simplex $\|x\|_1 = d\, y_1$ and the sphere $\|x\|_2^2 = d\, y_2$. Since $\|x\|_2^2 \leq \|x\|_1^2 \leq d\,\|x\|_2^2$, this intersection is non-empty over a wider range defined by

$$1 \leq \frac{y_2}{y_1^2} \leq d .$$

When $1 < y_2/y_1^2 \leq 2$, micro and macrocanonical densities have the same limit when $d$ goes to $\infty$ and $\epsilon$ goes to zero. S. Chatterjee [13] proves that the joint microcanonical density of $X_{d,\epsilon}(1), \ldots, X_{d,\epsilon}(m)$ for a fixed $m$ converges to i.i.d random variables having an exponential distribution equal to $\alpha e^{-\beta_1 |z| - \beta_2 |z|^2}$, and $(\mathbb{E}(|X_{d,\epsilon}(u)|^1, \mathbb{E}(|X_{d,\epsilon}(u)|^2)$ converges to $y$. If $y_2/y_1^2 = 2$ then $\beta_2 = 0$. In this regime where macrocanonical densities are well-defined, micro and macrocanonical measures converge to each other so the Boltzmann equivalence principle is again verified.

However, when $y_2/y_1^2 > 2$ the macrocanonical density is not defined, so the Boltzmann equivalence principle is violated. The microcanonical set contains sparse signals which are not captured by exponential distributions. In this case, Chatterjee [13] proves that when $d$ goes to $\infty$ and $\epsilon$ to 0, $X_{d,\epsilon}$ has one large coefficient randomly located at some $u_0 \in \Lambda_d$ for which $X_{d,\epsilon}^2(u_0) \sim d(y_2 - 2y_1^2)$ with a probability which tends to 1. All other coefficients have a much smaller $O(y_1)$ amplitude. For $m$ fixed, $X_{d,\epsilon}(1), \ldots, X_{d,\epsilon}(m)$ converge in law to i.i.d random variables having marginals equal to $e^{-\beta_1 |z|}$, but there is no convergence of moments. This example shows that the Boltzmann equivalence principle is not necessarily satisfied, particularly when signals exhibit a strong sparsity behavior.

**2.4. Boltzmann equivalence principle.** Micro and macrocanonical densities are defined over configurations $x$ specified in a finite cube $\Lambda_d$ of dimension $\ell$. Let $\Phi_d(x)$ be a shift-invariant energy vector computed by averaging a finite range potential $Ux$. To compute estimators which converge when $d$ goes to $\infty$, we need to ensure that microcanonical densities converge in the moments sense. We consider the limit among measures defined on the configuration space $\mathcal{I}^{\mathbb{Z}^\ell}$, with the product topology of Borel fields on the interval $\mathcal{I} \subset \mathbb{R}$. The asymptotic equivalence between micro and macrocanonical measures is called the Boltzmann equivalence principle [22]. Their convergence to the same Gibbs measures was first proved by Landford [30]. It is the center of a large body of work, rooted in the theory of large deviations [20]. We review results obtained when $\mathcal{I}$ is a bounded interval and for Gaussian processes.

**Macrocanonical convergence.** When $\mathcal{I}$ is a bounded interval, macrocanonical distributions are unique minimizers of the Lagrangian (7). When $d$ goes to $\infty$, the limit Gibbs measure is defined by normalizing this Lagrangian so that it converges to a variational problem defined over a stationary measure $\mu$. Suppose that $\mu$ exists. Since $Ux$ is equivariant to translations and $\mu$ is stationary it results that $\mathbb{E}_\mu(Ux(u)) = \mathbb{E}_\mu(Ux)$ does not depend upon the grid point $u$. Suppose that $\mu$ has no long range correlation so that boundary values have a negligible influence. Since $\Phi_d(x)$ is an average of $Ux(u)$ in $\Lambda_d$ it follows that

$$\lim_{d\to\infty} \mathbb{E}_\mu(\Phi_d(x)) = \mathbb{E}_\mu(Ux)\,.$$

The Lagrangian (7) includes a negative entropy term that diverges as $d \to \infty$ if $\mu$ has finite range correlations. The normalisation replaces the entropy by an entropy rate $\bar{H}(\mu)$, defined by considering the restriction $\mu_d$ of $\mu$ on the finite dimensional configuration space $\mathcal{I}_d^\Lambda$. Let $q_d$ be the density of $\mu_d$ relatively to the Lebesgue measure. If $\mu$ has a finite range correlation we expect that $H(q_d)$ grows linearly with $d$. The entropy rate is defined by

$$\bar{H}(\mu) = \lim_{d\to\infty} d^{-1}\, H(q_d)\,. \tag{15}$$

Normalizing the free energy Lagrangian (7) by $d$ and taking the limit when $d$ goes to $\infty$ defines a new Lagrangian

$$\mathcal{L}_\infty(\mu,\beta) = -\bar{H}(\mu) + \langle \beta, \mathbb{E}_\mu(Ux) - y\rangle\,. \tag{16}$$

Gibbs measures minimize this Lagrangian over the space of stationary measures for $\beta$ fixed.

If $U$ is a bounded, finite range and continuous potential, then one can prove [17,25] that the set of Gibbs measures which minimize this Lagrangian is a non-empty, convex and compact set of measures. In general the solution is not unique because contrarily to the finite Lagrangian (7) where $-H(p)$ is strictly convex, the entropy rate $\bar{H}(\mu)$ is affine [17, 25]. This implies that depending upon boundary conditions in $\Lambda_d$, macrocanonical densities may converge to different Gibbs measures, which is a phase transition phenomena.

Periodic boundary conditions over the finite cube $\Lambda_d$ simplify computational algorithms, but they are artificial. The limit Gibbs measure will not depend upon these boundary conditions if it is unique, and hence if there is no phase transition. This happens when there is no long range interactions, so that boundary values do not condition the probability distributions of far away values. In this paper, we concentrate on problems where there is no such phase transition.

**Microcanonical convergence.** The main difficulty is to find conditions which guarantee that microcanonical measures converge to the same Gibbs measure, having a maximum entropy rate conditioned by moment conditions. Suppose that $U$ is continuous, bounded and has a finite range. When $d$ goes to $\infty$ and $\epsilon$ goes to zero, one can prove [17, 25] that microcanonical distributions converge for an appropriate topology, to a limit measure which minimizes the same Lagrangian (16) as the one obtained from macrocanonical densities. If there is no phase transition, so that the macrocanonical measure converges to a unique Gibbs measure $\mu$, then this limit is the same for macrocanonical and microcanonical measures. More specifically, if $f(x)$ is a bounded and continuous function defined for any $x \in \mathcal{I}^{\mathbb{Z}^\ell}$, then the expected value of $f$ computed over $\Lambda_d$ with microcanonical and macrocanonical measures converge to $\mathbb{E}_\mu(f(x))$ when $d$ goes to $\infty$. We thus have a convergence for all bounded moments. However, it is not necessary to impose that $\mathcal{I}$ is bounded to verify the Boltzmann equivalence principle, as shown by the following Gaussian example.

**Gaussian processes.** Gaussian stationary measures are important examples of Gibbs measures where $x$ takes its values in $\mathcal{I} = \mathbb{R}$. They are obtained with a quadratic potential $Ux = \{U_k x\}_{k \leq K}$ computed with convolutions so that it is equivariant to translations over the grid $\mathbb{Z}^d$. Let us define

$$U_k x(u) = |x \star h_k(u)|^2 = \Big| \sum_{m \in \mathbb{Z}^d} x(u-m)\, h_k(m) \Big|^2 ,$$

where each $h_k$ has a support in $[-\Delta, \Delta]$.

If $x \in \mathbb{R}_d^\Lambda$ then $Ux$ is computed by extending $x$ on $\mathbb{Z}^\ell$ with a periodic extension beyond boundaries. Potentials can then be rewritten with circular convolutions of $x$

$$U_k x(u) = |x \star h_{d,k}(u)|^2 = \Big| \sum_{m \in \Lambda_d} x(m)\, h_{d,k}(n-m) \Big|^2 . \tag{17}$$

with periodic filters

$$h_{d,k}(n) = \sum_{m \in \mathbb{Z}^\ell} h_k\big(n - m d^{1/\ell}\big) . \tag{18}$$

The energy $\Phi_d(x)$ is thus a vector of normalized $\mathbf{l}^2$ norms:

$$\Phi_d(x) = \Big\{ d^{-1} \sum_{u \in \Lambda_d} |x \star h_{d,k}(u)|^2 = d^{-1} \|x \star h_{d,k}\|_2^2 \Big\}_{k \leq K} . \tag{19}$$

If $\widehat{h}_k(\omega)$ does not vanish for all $\omega \in [0, 2\pi]$ and $k \leq K$ then Varadhan and Donsker [19] proved that Boltzmann equivalence principle is satisfied when $d$ goes to $\infty$. The microcanonical and macrocanonical models converge to a Gaussian stationary process $\mu$ whose power-spectrum is

$$P_\mu(\omega) = \Big( \sum_{k=1}^{K} \beta_k |\widehat{h}_k(\omega)|^2 \Big)^{-1} . \tag{20}$$

The next section studies asymptotic properties of microcanonical models even though the macrocanonical model may not exist.

## 3. Microcanonical models beyond Boltzmann equivalence

We can guarantee that a maximum entropy microcanonical measure exists by making sure that microcanonical ensembles are compact. Even if this valid, the macrocanonical measure may not exist if $x(u)$ is defined over an interval $\mathcal{I}$ which is not bounded. In this case the Boltzmann equivalence principle is violated. Section 2.3 gives an example with uniform measures over intersections of $\mathbf{l}^1$ and $\mathbf{l}^2$ balls, in the sparse regime. Microcanonical models thus offer more flexibility, particularly for signals having sparse behavior.

In the rest of the paper, we embed all processes over $\mathbb{R}$, including binary processes such as Ising and Bernoulli . We thus consider that $x(u)$ takes its values in $\mathcal{I} = \mathbb{R}$, in which case $\mathcal{I}_d^\Lambda = \mathbb{R}^d$, where the grid topology is omitted for ease of notation. We study microcanonical properties independently from the corresponding macrocanonical measures which may not exist. For this purpose, Section 3.1 relates the maximum entropy of a microcanonical measure to the Jacobian of the energy potential. It gives sufficient conditions so that the entropy rate converges when $d$ goes to $\infty$. However, sampling a maximum entropy microcanonical process is computationally very expensive. Section 3.2 introduces a different class of microcanonical processes obtained by transporting a maximum entropy measure with a gradient descent algorithm which converges towards the microcanonical set. The transported measure does not have a maximum entropy but we prove that it has common symmetries with the maximum entropy measure. Convergence to microcanonical sets is studied in Section 3.3.

**3.1. Microcanonical entropy and Jacobian.** We study the convergence of maximum entropy microcanonical models when $d$ goes to $\infty$ by studying the convergence of their entropy rate without supposing that there exists a macrocanonical model. This is done by relating the maximum entropy rate to the Jacobian of the energy $\Phi_d$.

We consider a shift-equivariant and finite range potential from Section 2.3, and the corresponding microcanonical measure $\mu^{\mathrm{mi}}_{d,\epsilon}$, defined as the uniform distribution on compact sets of the form

$$\Omega_{d,\epsilon} = \left\{x \in \mathbb{R}^d : \|\Phi_d(x) - y\| \le \epsilon\right\}.$$

We saw in (5) that the entropy of $\mu^{\mathrm{mi}}_{d,\epsilon}$ is

$$H(\mu^{\mathrm{mi}}_{d,\epsilon}) = -\int p_{d,\epsilon}(x)\, \log p_{d,\epsilon}(x)\, dx = \log\left(\int 1_{\Omega_{d,\epsilon}}(x)\, dx\right). \tag{21}$$

Since $\Phi_d(x) = d^{-1} \sum_{u \in \Lambda_d} Ux(u)$ and $Ux(u)$ only depends on the values of $x(i)$ for $i \in [u - \Delta, u + \Delta]^\ell$, one can verify that the $i$-th column $J_i \Phi_d(x) = \partial_{x(i)} \Phi_d(x) \in \mathbb{R}^K$ only depends upon the restriction of $x$ in $[i - \Delta, i + \Delta]^\ell$. Moreover, thanks to the equivariant structure of $U$, one can verify that

$$\forall\, i \leq d \ , \ J_i \Phi_d(x) = d^{-1} \sum_{|m| \leq \Delta} \partial_{x(0)} U(T_{-i}) x(m) \,,$$

so the global properties of the Jacobian $J\Phi_d(x)$ can be derived from the Jacobian of the potential, restricted on a window:

$$\begin{aligned} JU \colon \mathbb{R}^{(2\Delta+1)\ell} &\to \mathbb{R}^K \\ x &\mapsto \sum_{|m| \leq \Delta} \partial_{x(0)} Ux(m) \,. \end{aligned} \tag{22}$$

We denote by $\partial \mathcal{A}$ the frontier of a set $\mathcal{A}$ and by $\mathcal{A}^o = \mathcal{A} - \partial \mathcal{A}$ the interior of $\mathcal{A}$, and by $\bar{\mathcal{A}}$ the complement of $\mathcal{A}$. We also denote by $|J\Phi_d(x)| = \sqrt{\det(J\Phi_d(x) J\Phi_d(x)^T)}$ the $K$-dimensional determinant of $J\Phi_d$, and by $d\mathcal{H}(x)^L$ the $L$-dimensional Hausdorff measure. We shall make the following assumptions on $U$:

(A) $U$ is uniformly Lipschitz on compact sets, which implies that for any compact $\mathcal{C} \subset \mathbb{R}^d$ there exists $\beta \geq 0$ such that

$$\forall (x, x') \in \mathcal{C}^{2d} \ , \ \|\Phi_d(x) - \Phi_d(x')\|_2 \leq \beta \, \|x - x'\|_2 \,. \tag{23}$$

It also implies that $|J\Phi_d(x)| \leq \beta^K$ for $x \in \mathcal{C}$.

(B) We shall also suppose that $\Phi_d^{-1}$ maps compact sets $\mathcal{C}$ to compact sets, with a controlled growth with respect to $d$. For each compact set $\mathcal{C} \subset \mathbb{R}^K$, there exists a constant $C$ independent of $d$ such that

$$\forall\, d \ , \ \Phi_d^{-1}(\mathcal{C}) = \left\{ x \in \mathbb{R}^d : \Phi_d(x) \in \mathcal{C} \right\} \subseteq B_{2,d}(C\sqrt{d}) \,, \tag{24}$$

where $B_{p,d}(R)$ denotes the $d$-dimensional $\mathbf{l^p}$ Euclidean Ball of radius $R$. It follows that $\Phi_d^{-1}(y)$ is a compact and Lipschitz manifold whose dimension is typically $d - K$, except for degenerated cases. For example, if $d^{-1}\|x\|_2^2$ is a component of the vector $\Phi_d$, this condition is satisfied.

Lastly, we need to control the integrability of $|J\Phi_d|^{-1}$ nearby microcanonical sets. More precisely, for each $y$ and any sufficiently small $\epsilon > 0$, we require that $|J\Phi_d(x)|^{-1}$ is integrable in $\Omega_{d,\epsilon}^y$. The following gives a sufficient condition which depends only on the potential function.

(C) For some $R > 0$, let $X$ be drawn from the uniform measure in the ball $B(2\Delta + 1, R)$ and $Z = JU(X) \in \mathbb{R}^K$ be the random vector obtained by applying the mapping $JU$ defined in (22). We shall suppose that there exists $\eta > 0$ such that

$$\forall\, \mathcal{S} \subset \mathbb{R}^K \text{ Lebesgue measurable} \ , \ P(Z \in \mathcal{S}) \lesssim |\mathcal{S}|^\eta \,. \tag{25}$$

This condition assumes that the differential of $U$ does not concentrate too much on a low-dimensional subspace of $\mathbb{R}^K$, nor in a discrete subset, but it does not require that its distribution is absolutely continuous with respect to the Lebesgue measure. We shall see next that potentials of the form $Ux = \{|x \star h_k|^p\}_{k \le K}$ with $p = 1, 2$ with complex filters $h_k$ define an integrable $|J\Phi_d|^{-1}$.

The following theorem computes the entropy of a microcanonical process from a change of variable metric, which depends upon the Jacobian of the interaction energy $\Phi_d$. The theorem derives a microcanonical entropy rate which converges when $d$ goes to $\infty$.

**Theorem 3.1.** *Suppose $U$ verifies (A), (B), and (C) above. Then the following properties are verified:*

(i) *For sufficiently large $d$,*

$$H(\mu^{\mathrm{mi}}_{d,\epsilon}) = \log \int_{\|z-y\| \le \epsilon} \gamma_d(z)\, dz\,, \tag{26}$$

*where $\gamma_d$ is the change of variable metric which satisfies*

$$\gamma_d(y) = \int_{\Phi_d^{-1}(y)} |J\Phi_d(x)|^{-1}\, d\mathcal{H}^{d-K}(x) < \infty \ a.e., \tag{27}$$

*where $\mathcal{H}^{d-K}$ is the $d-K$ dimensional Hausdorff measure. Moreover, $\gamma_d(y)$ has a finite integral on compact sets.*

(ii) *The function $\gamma_d$ is strictly positive in the interior of $\Phi_d(\mathbb{R}^d)$, up to a thin shell on the boundary; i.e. on sets $C_d \subset \Phi_d(\mathbb{R}^d)$ satisfying*

$$\sup_{y \in C_d} \mathrm{dist}(y, \overline{\Phi_d(\mathbb{R}^d)}) \le c \cdot d^{-1/\ell}\,,$$

*for some constant $c$.*

(iii) *Suppose that either $\Delta = 1$, or that the potential $U$ is Hölder continuous with parameter $\alpha < 2/\ell$: $|U(x) - U(x')| \le C\|x - x'\|^\alpha$. Then, for each $\epsilon > 0$, the entropy rate $d^{-1}H(\mu^{\mathrm{mi}}_{d,\epsilon})$ converges as $d \to \infty$ and satisfies*

$$-\infty < \lim_{d \to \infty} d^{-1} H(\mu^{\mathrm{mi}}_{d,\epsilon}) \le C \log \|y\|^2\,, \tag{28}$$

*where $C$ is a universal constant.*

The proof is in Appendix A. This theorem highlights the connection between the entropy and the Jacobian through $\gamma_d(y)$, via the coarea formula. It defines the entropy rate of a microcanonical ensemble for general $\Phi_d$ in the thermodynamical limit $d \to \infty$, without relying on a macrocanonical model. One can compare the conditions of Theorem 3.1 with those that ensure the convergence of the microcanonical and macrocanonical measures. In [16, 43] this equivalence is established for bounded, finite-range potentials $U$. Our condition to prove that the entropy rate converges is weaker ($U$ Hölder continuous), but we do not study convergence beyond the entropy rate. Studying the convergence of the microcanonical measure in more general conditions remains an open question. Finally, notice that for positive integers $k$, the Hausdorff measure is equivalent to the $k$-dimensional Lebesgue measure up to a constant rescaling.

The microcanonical thickness parameter $\epsilon$ is important to ensure appropriate convergence. The following corollary quantifies the effect of $\epsilon$ in the entropy rate, and proves that its contribution to the energy is small for sufficiently large $d$.

**Corollary 3.2.** *Under the same conditions as Theorem 3.1, for $d$ fixed and when $\epsilon \to 0$, the entropy rate of the $\epsilon$-thick microcanonical model satisfies*

$$d^{-1} H(\mu^{\rm mi}_{d,\epsilon}) \sim \frac{K}{d} \log \epsilon.$$

As a consequence of this corollary, the entropy variation due to a change in the thickness from $\epsilon$ to $\epsilon'$ is of the order of $\frac{K}{d} \log\left(\frac{\epsilon}{\epsilon'}\right)$, which is negligible if $K \log\left(\frac{\epsilon}{\epsilon'}\right) \ll d$.

This paper concentrates on interaction energy vectors $\Phi_d$ defined by $\mathbf{l}^2$ and $\mathbf{l}^1$ norms of convolutions of $x$ with multiple filters. The next proposition proves that such interaction energies satisfy the assumptions of Theorem 3.1. The proof is in Appendix C.

**Proposition 3.3.** *$\Phi_d$ satisfies assumptions (A), (B), and (C) in the following cases:*

(i) $\Phi_d(x) = \{d^{-1}\|x \star h_k\|_2^2\}_{k \le K}$ *and the $\{h_k\}_{k\le K}$ are linearly independent.*

(ii) $\Phi_d(x) = \{d^{-1}\|x\|^2, d^{-1}\|x\|_1\}$.

(iii) $\Phi_d(x) = \{d^{-1}\|x\|^2, d^{-1}\|x \star h_k\|_1\}_{k\le K}$ *and the $h_k$ are linearly independent with $|\widehat{h}_k(-\omega)| \neq |\widehat{h}_k(\omega)|$ for all $\omega$.*

**3.2. Microcanonical gradient descent model.** Computing samples of a maximum entropy microcanonical model is typically done with MCMC algorithms or Langevin Dynamics [15], which is computationally very expensive. Computations can be considerably reduced by avoiding to enforce the maximum entropy constraint over the microcanonical set. Microcanonical models computed with alternative projections and gradient descents have been implemented to sample texture synthesis models [23, 26, 39]. Another related sampling algorithm is the so-called Herding algorithm by Welling [47], which produces "pseudo-samples" of a microcanonical model in a deterministic fashion by solving a sequence of primal-dual updates.

We consider microcanonical gradient descent models obtained by transporting an initial measure towards a microcanonical set, using gradient descent with respect to the distance to the microcanincal ensemble. We prove that the gradient descent preserves many symmetries of the maximum entropy microcanonical measure.

Let $\Phi_d$ be a shift-invariant function as defined in Section 2.3 and $y \in \Phi_d(\mathbb{R}^d)$. We transport an initial measure $\mu_0$ towards a measure supported in a microcanonical set $\Omega_{d,\epsilon}$, by iteratively minimising

$$E(x) = \frac{1}{2}\|\Phi_d(x) - y\|^2 \tag{29}$$

with mappings of the form

$$\varphi_n(x) = x - \kappa_n \nabla E(x) = x - \kappa_n J\Phi_d(x)^T(\Phi_d(x) - y)\,, \tag{30}$$

where $\kappa_n$ is the gradient step at each iteration $n$.

Given an initial measure $\mu_0$, the measure update is

$$\mu_{n+1} := \varphi_{n,\#}\mu_n, \tag{31}$$

with the standard pushforward measure $f_\#(\mu)[\mathcal{A}] = \mu[f^{-1}(\mathcal{A})]$ for any $\mu$-measurable set $\mathcal{A}$, where $f^{-1}(\mathcal{A}) = \{x; f(x) \in \mathcal{A}\}$.

Samples from $\mu_n$ are thus obtained by transforming samples $x_0$ from $\mu_0$ with the mapping $\overline{\varphi} = \varphi_n \circ \varphi_{n-1} \circ \cdots \circ \varphi_1$. It corresponds to $n$ steps of a gradient descent initialized with $x_0 \sim \mu_0$:

$$x_{l+1} = x_l - \kappa_l J\Phi_d(x_l)^T(\Phi_d(x_l) - y)\,.$$

Next section studies the convergence of the gradient descent measures $\mu_n$. Even if they converge to a measure supported in a microcanonical set $\Omega_{d,\epsilon}$, in general they do not converge to a maximum entropy measure on this set. However, the next theorem proves that if $\mu_0$ is a Gaussian measure of i.i.d Gaussian random variables then they have a large class of common symmetries with the maximum entropy measure. Let us recall that a symmetry of a measure $\mu$ is a linear invertible operator $L$ such that for any measurable set $\mathcal{A}$, $\mu[L^{-1}(\mathcal{A})] = \mu[\mathcal{A}]$. A linear invertible operator $L$ is a symmetry of $\Phi_d$ if for all $x \in \mathbb{R}^d$, $\Phi_d(L^{-1}x) = \Phi_d(x)$. It preserves volumes if its determinant satisfies $|\det L| = 1$. It is orthogonal if $L^tL = LL^t = I$ and we say that it preserves a stationary mean if $L\mathbf{1} = \mathbf{1}$ for $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^\ell$.

**Theorem 3.4.** (i) *If $L$ is a symmetry of $\Phi_d$ which preserves volumes then it is a symmetry of the maximum entropy microcanonical measure.*

(ii) *If $L$ is a symmetry of $\Phi_d$ and of $\mu_0$ then it is a symmetry of $\mu_n$ for any $n \geq 0$.*

(iii) *Suppose that $\mu_0$ is a Gaussian white noise measure of $d$ i.i.d Gaussian random variables. Then, if $L$ is a symmetry of $\Phi_d$ which is orthogonal and preserves a stationary mean then it is a symmetry of $\mu_n$ for any $n \geq 0$.*

The theorem proof is in Appendix D. The initial measure $\mu_0$ is chosen so that it has many symmetries in common with $\Phi_d$ and hence the gradient descent measures have many symmetries in common with a maximum entropy measure. A Gaussian measure of i.i.d Gaussian variables of mean $m_0$ and $\sigma_0$ is a maximum entropy measure conditioned by a stationary mean and variance. It is uniform over spheres which guarantees that it has a large group of symmetries. The stationary mean $m_0$ and variance $\sigma_0^2$ are adjusted so that that microcanonical sets are nearly included over the sphere of mean $m_0\mathbf{1}$ and radius $\sigma_0$, where $\mu_0$ concentrates and is uniform. We thus set $m_0$ and $\sigma_0^2$ to be the empirical stationary mean and variance calculated from the realization $\bar{x}$ of $X$:

$$m_0 = d^{-1} \sum_{u \in \Lambda_d} \bar{x}(u) \quad \text{and} \quad \sigma_0^2 = d^{-1} \sum_{u \in \Lambda_d} (\bar{x}(u) - m_0)^2 \,. \tag{32}$$

Observe that periodic shifts are linear orthogonal operators and preserve a stationary mean. The following corollary applies property (iii) of Theorem 3.4 to prove that $\mu_n$ are circular-stationary.

**Corollary 3.5.** *If $\Phi_d$ is invariant to periodic shift and $\mu_0$ is a Gaussian white noise then $\mu_n$ is circular-stationary for $n \geq 0$.*

**3.3. Convergence of microcanonical gradient descent.** This section studies conditions so that the gradient descent (31) converges to a stationary measure supported in a microcanonical ensemble, and we give a lower bound of its entropy rate. To guarantee that the algorithm is not trapped in local minima, we use the characterization of stable solutions from [31,38] based on the second-order analysis of critical points of (29). Such analysis reveals that gradient descent methods do not get stuck at critical points which are *strict saddles* — in which at least one Hessian eigenvalue is strictly negative, since the set of initialization parameters corresponding to the non-negative spectrum has measure 0 relative to $\mu_0$.

**Definition 3.6.** We say that $\Phi_d = (\phi_1, \dots, \phi_K)$ has the strict saddle condition if $\Phi_d$ is at least $\mathbf{C}^2$ and for each $v \in \mathrm{Null}(J\Phi_d(x)^\top) \subseteq \mathbb{R}^K$, $v \neq 0$, the matrix

$$\sum_{k \leq K} v_k \nabla^2 \phi_k(x) + J\Phi_d(x)^\top J\Phi_d(x) \tag{33}$$

has at least one strictly negative eigenvalue, where $\nabla^2\phi_k$ is the Hessian of $\phi_k$.

The following theorem, proved in Appendix E, establishes basic properties of the distribution generated by gradient descent, including sufficient conditions for its convergence to the microcanonical ensemble.

**Theorem 3.7.** *Assume $\Phi_d$ is $\mathbf{C}^2$ and satisfies property (B) (24). Suppose that $\Phi_d$ is Lipschitz with $Lip_{\Phi_d} = \beta$ and that $\nabla\Phi_d$ is also Lipschitz, with $Lip_{\nabla\Phi_d} = \eta$. Let $y \in \Phi_d(\mathbb{R}^d)^\circ$. Then:*

(i) *If $\Phi_d$ satisfies the strict saddle condition, then (29) has no poor local minima. Moreover, if $|J\Phi_d(x)| > 0$ for all $x \in \Phi_d^{-1}(y)$, then by choosing step-sizes $\kappa_n$ such that $\kappa_n < \eta^{-1}$ for all $n$, $\mu_n$ converges almost surely to a limit measure $\mu_\infty$ [1] Moreover, $\mu_\infty$ is supported in the microcanonical ensemble $\Phi_d^{-1}(y)$ with appropriate choice of learning rate $\kappa_n$; that is, $A \cap \Phi_d^{-1}(y) = \emptyset \Rightarrow \mu_\infty(A) = 0$.*

(ii) *The entropy rate $d^{-1}H(\mu_n)$ satisfies*

$$d^{-1}H(\mu_n) \geq d^{-1}H(\mu_0) - \left(1 - \frac{K}{d}\right)\eta \sum_{n' \leq n} \kappa_{n'} r_{n'} - \frac{K}{d}\beta^2 \sum_{n' \leq n} \kappa_{n'}\,, \quad (34)$$

*where $r_n = \mathbb{E}_{\mu_n}\sqrt{E(x)}$ is the average distance to the microcanonical ensemble at iteration $n$.*

Part (i) gives sufficient conditions for the gradient descent sampling to converge towards the microcanonical ensemble. Each gradient descent step can reduce the entropy rate. By computing an upper bound of this entropy reduction, part (ii) gives a lower bound of the entropy rate after $n$ iterations. Although the gradient descent converges to the microcanonical ensemble in general the resulting measure will not have a maximum entropy. However, (34) gives a lower bound of its entropy rate. By choosing a measure $\mu_0$ having a maximum entropy, we maximize the entropy of the lower-bound (34).

Our current results rely on second-order stationarity assumptions, but first-order stationary condition $\nabla E(x^*) = 0$ may be sufficient to characterize convergence as $d \to \infty$. Indeed, this condition implies that either we reached the microcanonical ensemble, $\Phi(x^*) = y$, or that we have found a non-regular point, with $|J\Phi_d(x^*)| = 0$. Such points occur with vanishing probability as $d \to \infty$, but the rigorous analysis of this phenomena is left for future work.

The sufficient condition for $\mu_n$ to converge to a limit measure $\mu_\infty$ requires $|J\Phi_d(x)| > 0$ for $x \in \Phi_d^{-1}(y)$, which for certain choices of $\Phi_d$ may be hard to check. The following corollary, proved in Appendix F, provides an alternative sufficient condition which is stronger but easier to evaluate.

**Corollary 3.8.** *If $\Phi_d$ is $\mathbf{C}^\infty$ and Lipschitz and satisfies the strict saddle condition, then $\mu_n$ converges for any $y \in \Phi_d(\mathbb{R}^d)$ up to a set of zero measure, and $\mu_\infty$ is supported in the microcanonical ensemble.*

We now give examples of energies $\Phi_d$ which satisfy the assumptions of previous theorem. The next theorem, proved in Appendix G, shows that the $\mathbf{l}^2$

[1] Defined as $\mathrm{Prob}[\mu_n(A) \to \mu_\infty(A)$ for any $\mathcal{F}$-measurable set $A] = 1$, where $\mathcal{F}$ is the Borel $\sigma$-algebra on $\mathbb{R}^d$.

ellipsoid representation satisfies the strict saddle condition, and therefore that the microcanonical gradient descent measure is supported in the microcanonical ensemble.

**Theorem 3.9.** *If $\Phi_d(x) = \{d^{-1}\|x \star h_k\|_2^2\}_k$ with linearly independent and compactly supported $h_k$, then $\Phi_d$ satisfies the strict saddle condition and $|J\Phi_d(x)| > 0$ for $x \in \Phi_d^{-1}(y)$ with $y \in \Phi_d(\mathbb{R}^d)^\circ$, and therefore $\mu_\infty$ is supported in the microcanonical ensemble.*

A current limitation of the convergence analysis is that it relies on smoothness properties of $\Phi_d$, thus leaving out of scope the $\mathbf{l}^1$-based representations. This limitation is intrinsic to the convergence analysis of non-smooth, non-convex optimization methods, which provides no guarantees using simple gradient descent. The analysis of other algorithms such as ADMM [46] or gradient sampling [12] in such conditions is left for future work.

**Continuous-time limit dynamics.** The measure transport (31) defined by gradient descent can be seen as a discretization of an underlying partial differential equation in the space of measures, describing the behavior as the step-size $\kappa_n \to 0$. The resulting dynamics is described by the well-known *continuity equation*, expressed in the distributional sense as

$$\partial_t \mu_t = \operatorname{div}(\nabla E \cdot \mu_t)\,, \tag{35}$$

or equivalently

$$\forall\, \phi \in \mathbf{C}^1\ ,\ \partial_t\Big(\int \phi(x)\mu_t(dx)\Big) = -\int \langle \nabla\phi(x), \nabla E(x)\rangle \mu_t(dx)\,,$$

where $\mathbf{C}_c^1$ denotes the space of $\mathbf{C}^1$ compactly supported test functions. As opposed to MCMC algorithms, which are discretizations of diffusion Stochastic Differential Equations (SDEs), the dynamics in our case are deterministic, and the only source of randomness comes from the initial measure $\mu_0$. Notice also that the symmetry preservation properties described in Theorem 3.4 directly apply to the Liouville equation above. Equation (35) can also be interpreted as a Wasserstein Gradient Flow over the functional energy

$$\mathcal{E}[\mu] = \int E(x)\mu(dx)\,.$$

Recent work [14,40] has established global convergence of such Wasserstein Gradient Flows in the cases where $E$ is positively homogeneous, for suitable initialization. Although in our case $E$ is not homogeneous, we leave for future work to exploit the homogeneity properties of $\Phi_d$ to derive similar convergence results that can generalize Theorem 3.7.

## 4. Multiscale microcanonical wavelet and scattering models

We study multiscale microcanonical models obtained with energy vectors computed with a wavelet transform. Next section introduces energy vectors computed with $\mathbf{l}^2$ and $\mathbf{l}^1$ norms of wavelet coefficients. Section 4.2 introduces scattering which provide complementary $\mathbf{l}^1$ norm coefficient computed with a second wavelet transform.

**4.1. Wavelet transform $\mathbf{l}^2$ and $\mathbf{l}^1$ norms.** A wavelet transform, computes signal variations at different scales through convolutions with dilated wavelets. Maximum entropy models conditioned by wavelet $\mathbf{l}^2$ norms define Gaussian processes. Wavelet transforms define sparse representations of large classes of signals. This sparsity characterize non-Gaussian behavior which is specified by wavelet $\mathbf{l}^1$ norms. We write $\widehat{x}$ the Fourier transform of $x$.

**Wavelet transform.** Wavelet coefficients are convolutions $x \star \psi_{j,q}(u)$ for $u \in \mathbb{R}^\ell$, where each wavelet $\psi_{j,q}$ is a dilated band-pass filter which covers different frequency domains:

$$\psi_{j,q}(u) = 2^{-\ell j}\psi_q(2^{-j}u) \;\Rightarrow\; \widehat{\psi}_{j,q}(\omega) = \widehat{\psi}_q(2^j\omega). \tag{36}$$

We will focus our attention on the compactly-supported case, where the $Q$ mother wavelets $\psi_q$ have a support in $[-C, C]^\ell$ so the support of $\psi_{j,q}$ is in $[-C2^j, C2^j]^\ell$. The Fourier transform $\widehat{\psi}_q(\omega)$ have an energy concentrated in frequency intervals which barely overlap for different $q$.

If $x$ is supported in a cube $\Lambda_d \subset \mathbb{Z}^\ell$, then $u$ is discretized on this square grid. Convolutions are defined by extending $x$ into a periodic signal over $\mathbb{Z}^\ell$. We showed in (17) that it is equivalent to computing circular convolutions with periodic wavelet filters (18). Discrete periodic wavelets $\psi_{j,q}$ are band-pass filters with a zero average $\sum_{u\in\Lambda_d}\psi_{j,q}(u) = 0$. The minimum scale $2^j$ is limited by the sampling interval normalized to 1, whereas the maximum scale $2^J$ is limited by the width $d^{1/\ell}$ of $\Lambda_d$.

Wavelet coefficients $x \star \psi_{j,q}(u)$ separate the frequency components of $x$ in several frequency bands, at scales $1 \leq 2^j \leq 2^J$. The remaining low frequencies at scales larger than $2^J$ are carried by a single low-pass filter which we write $\psi_{J,0}(u) = 2^{-Jd}\psi_0(2^{-J}u)$, whose support is also included in $[-C2^J, C2^J]^\ell$.

The wavelet transform of $x$ is defined by

$$Wx = \left\{x \star \psi_{j,q}\right\}_{1\leq j\leq J, q\leq Q}. \tag{37}$$

We impose that the frequency supports $\widehat{\psi}_{j,q}$ cover uniformly the whole frequency domain, which is captured by the following Littlewood–Paley condition. There exists $\gamma < 1$ such that

$$\forall \omega\,,\; 1-\gamma \leq |\widehat{\psi}_{J,0}(\omega)|^2 + \frac{1}{2}\sum_{j,q}\left(|\widehat{\psi}_{j,q}(\omega)|^2 + |\widehat{\psi}_{j,q}(-\omega)|^2\right) \leq 1+\gamma\,. \tag{38}$$

The condition implies the following energy inequalities for any $x \in I^{\Lambda_d}$

$$(1-\gamma)\|x\|_2^2 \leq \|x \star \psi_{J,0}\|_2^2 + \sum_{j,q} \|x \star \psi_{j,q}\|_2^2 \leq (1+\gamma)\,\|x\|_2^2\,. \tag{39}$$

This is proved by multiplying (38) with $|\widehat{x}(\omega)|^2$ and applying the Plancherel equality. This property implies that $W$ is a bounded and invertible operator, and its inverse has a norm smaller than $(1-\gamma)^{-1/2}$. If $\gamma = 0$ then $W$ is an isometry.

For audio signals in dimension $\ell = 1$, each wavelet is a complex filter whose Fourier transform $\widehat{\psi}_q(\omega)$ has an energy concentrated in the interval $[2^{q/Q}, 2^{(q+1)/Q}]$. It follows that $\widehat{\psi}_{j,q}(\omega)$ covers the interval $[2^{-j+q/Q}, 2^{-j+(q+1)/Q}]$ and satisfies the Littlewood–Paley condition (38). The parameter $Q$ is the number of wavelets per octave, which adjusts their frequency resolution. Wavelet representations are usually computed with about $Q = 12$ wavelets per octave, which are similar to half-tone musical notes. In numerical computations, we choose Gabor wavelets as in [2]. Although strictly speaking this wavelet family does not have spatially compact support, the decay is exponential and has no practical effect.

For images in $\ell = 2$ dimensions, each wavelet is computed by rotating a single mother wavelet

$$\psi_{j,q}(u) = 2^{-\ell j}\,\psi(2^{-j} r_q^{-1} u) \;\Rightarrow\; \widehat{\psi}_{j,q}(\omega) = \widehat{\psi}(2^j r_q \omega), \tag{40}$$

where $r_q u$ is a rotation of $u \in \mathbb{R}^2$ by an angle $q\pi/Q$. We choose a complex mother wavelet $\psi(u)$ whose Fourier transform $\widehat{\psi}(\omega)$ is centered at a frequency $\xi$ over a frequency domain of radius approximately $|\xi|/2$. The support of each $\widehat{\psi}_{j,q}$ is dilated and rotated according to (40). Wavelet coefficients $x \star \psi_{j,q}$ thus compute variations of $x$ at scales $2^j$ along different directions. In numerical computations we use Morlet wavelets as in [10] with $Q = 8$ angles to satisfy the Littlewood–Paley condition (38). As in the case of audio, these wavelets have exponentially decaying spatial envelop.

**Wavelet $\mathbf{l}^2$ norms.** We saw in Section 2.4 that microcanonical maximum entropy measures conditioned by energy vectors (50) of $\mathbf{l}^2$ norms converge to Gaussian processes. We can define such energy vectors with wavelet $\mathbf{l}^2$ norms, with the quadratic potential

$$Ux = \left\{|x \star \psi_{j,q}|^2\right\}_{j \leq J, q \leq Q}\,. \tag{41}$$

Since each filter support is included in $[-C2^J, C2^J]^\ell$, this potential has a finite range $\Delta = C2^J$. When $x$ is defined over a cube $\Lambda_d$ then $Ux$ is computed by periodizing $x$ which is equivalent to periodizing the wavelet filters and replacing convolutions with circular convolutions, as shown in (17). To simplify notations, the periodized filters are still written $\psi_{j,q}$. According to (50) the energy over a cube $\Lambda_d$ is given by normalized $\mathbf{l}^2$ norms

$$\Phi_d(x) = \left\{d^{-1}\,\|x \star \psi_{j,q}\|_2^2\right\}_{j \leq J, q \leq Q}. \tag{42}$$

It measures the energy of $x$ in the different frequency bands covered by each $\widehat{\psi}_{j,q}$.

**Wavelet $\mathbf{l}^1$ norms for sparsity.** Non-Gaussian properties can be captured with statistics sensitive to sparsity, as observed in early works studying the statistics of natural images [41], and formalized on specific processes such as multifractals [11]. Suppose that $X \star \psi_{j,q}(u)$ has few large amplitude coefficients and a large proportion of negligible coefficients. For example, if $X(u)$ is piecewise regular then $X \star \psi_{j,q}(u)$ is negligible over domains where $X(u)$ is regular and it has a large amplitude near singularities and sharp variations. The marginal probability density of $X \star \psi_{j,q}(u)$ is then highly concentrated near 0. It is thus better approximated by a Laplacian rather than a Gaussian distribution. We saw in Section 2.3 that Laplacian distributions are maximum entropy distributions conditioned by first order moments. This suggests to estimate $\mathbb{E}_\mu(|x \star \psi_{j,q}(u)|)$ as opposed to $\mathbb{E}_\mu(|x \star \psi_{j,q}(u)|^2)$, with a normalized $\mathbf{l}^1$ norm

$$d^{-1}\|x \star \psi_{j,q}(u)\|_1 = d^{-1} \sum_{u \in \Lambda_d} |x \star \psi_{j,q}(u)| \,.$$

A wavelet $\mathbf{l}^1$ norm energy is defined by replacing the quadratic potential (41) by a modulus potential

$$Ux = \left\{|x \star \psi_{j,q}|\right\}_{j \le J, q \le Q}\,, \tag{43}$$

which also has a finite range $\Delta = C2^J$. The resulting energy over a cube $\Lambda_d$ is

$$\Phi_d(x) = \left\{d^{-1}\,\|x \star \psi_{j,q}\|_1\right\}_{j \le J, q \le Q}. \tag{44}$$

It captures the sparsity of wavelet coefficients for each scale and orientation.

**4.2. Scattering transform.** Wavelet $\mathbf{l}^1$ norm measure the sparsity of wavelet coefficients but do not specify the spatial distribution of large amplitude wavelet coefficients. Scattering transforms provide information about this geometry by computing interaction terms across scales, with an iterated wavelet transform. Their mathematical properties are described in [11, 33], and applications to image and audio classification are studied in [2, 10]. We review important properties needed to define microcanonical models, including the energy conservation allowing to recover wavelet $\mathbf{l}^2$ norms.

The mean of $x$ is estimated over a cube $u \in \Lambda_d$ by $d^{-1}\sum_{u\in\Lambda_d} x(u)$. The modulus of a wavelet coefficient $|x \star \psi_{j,q}(u)|$ measures the variation of $x$ around its mean, in a neighborhood of $u$ of size proportional to $2^j$. A normalized $\mathbf{l}^1$ norm is the average of $|x \star \psi_{j,q}(u)|$

$$d^{-1}\|x \star \psi_{j,q}\|_1 = d^{-1} \sum_{u \in \Lambda_d} |x \star \psi_{j,q}(u)| \,.$$

Similarly, we can capture the variability of $|x \star \psi_{j,q}(u)|$ around this mean by convolving $|x \star \psi_{j,q}(u)|$ with a new set of wavelets:

$$||x \star \psi_{j,q}| \star \psi_{j',q'}(u)|.$$

It measures the variations of $|x \star \psi_{j,q}(u)|$ in a neighborhood of size $2^{j'}$. We shall consider the second wavelet $\psi_{j',q'}$ is calculated from the same mother wavelet than $\psi_{j,q}$ but for different $j', q'$, although the second mother wavelet may be changed as in [2].

The maximum scales $2^j$ and $2^{j'}$ remain below a cut-off scale $2^J$ which specifies the maximum interaction range of the model. Incorporating first and second order coefficients defines a new potential which captures the multiscale variations of $x$ as well as interaction terms across scales:

$$Ux = \left\{x\,,\ |x \star \psi_{j,q}|\,,\ ||x \star \psi_{j,q}| \star \psi_{j',q'}|\right\}_{j,j' \le J, q,q' \le Q}. \tag{45}$$

The corresponding energy vector is

$$\Phi_d(x) = \left\{d^{-1} \sum_{u \in \Lambda_d} x(u)\,,\ d^{-1}\,\|x \star \psi_{j,q}\|_1\,,\ d^{-1}\,\||x \star \psi_{j,q}| \star \psi_{j',q'}\|_1\right\}_{\substack{1 \le j,j' \le J,\\ q,q' \le Q}}. \tag{46}$$

It includes $K = 1 + JQ + J^2Q^2$ coefficients.

The following proposition, shows that wavelet $\mathbf{l}^2$ norms can be closely approximated from $\mathbf{l}^1$ norm scattering coefficients. As a result, we will be able to approximate Gaussian process as well as non-Gaussian processes with a scattering energy vector. It is proved in Appendix H,

**Proposition 4.1.** *Suppose that the wavelets satisfy* (38) *with* $\gamma = 0$ *then for* $J = \log_2 d$

$$\|x \star \psi_{j,q}\|_2^2 = \|x \star \psi_{j,q}\|_1^2 + \sum_{j'=1}^{\log_2 d} \sum_{q'=1}^{Q} \||x \star \psi_{j,q}| \star \psi_{j',q'}\|_1^2$$
$$+ \sum_{j'=1}^{\log_2 d} \sum_{q'=1}^{Q} \sum_{j''=1}^{\log_2 d} \sum_{q''=1}^{Q} \|||x \star \psi_{j,q}| \star \psi_{j',q'}| \star \psi_{j'',q''}\|_2^2. \tag{47}$$

This proposition proves that $\mathbf{l}^2$ of wavelet coefficients are approximated by sums of first and second order scattering coefficients plus a third order term

$$\sum_{j',q',j'',q''} \|||x \star \psi_{j,q}| \star \psi_{j',q'}| \star \psi_{j'',q''}\|_2^2.$$

For most stationary process this third order term is much smaller than the first two and can be neglected [10]. The theorem hypothesis supposes that wavelets satisfy the Littlewood inequality (38) with $\gamma = 0$. If $\gamma$ is non-zero, it creates corrective terms proportional to $(1 - \gamma)^2$. Observe also that we set $J = \log_2 d$. In microcanonical models, $2^J$ is a fixed scale so that the number of scattering coefficients does not increase with $d$.

## 5. Approximations of stationary processes

We study approximation of probability measures associated with stationary processes $X(u)$, $u \in \mathbb{Z}^\ell$, taking its values in $\mathbb{R}$, with gradient descent microcanonical models calculated with shift-invariant energy vectors. We first concentrate on Gaussian, Ising and point processes whose properties are well understood mathematically. We then consider the synthesis of image and audio textures from a single example.

**5.1. Microcanonical approximation errors.** This section analyzes the approximation errors of a stationary process $X$ of probability measure $\mu$ by a gradient descent microcanonical model of measure $\mu_n$. The gradient descent is initialized with a Gaussian white measure $\mu_0$ whose mean and variance are defined in (32). Since the energy $\Phi_d$ is shift-invariant, Corollary 3.5 proves that the gradient descent measures $\mu_n$ are stationary.

**Concentration.** Section 2.1 explains that a microcanonical model is based on a concentration hypothesis, which needs to be verified. For almost all realization $x$ of $X$, $\Phi_d(x)$ should remain in a ball of radius $\epsilon_d$ which converges to zero when $d$ goes to $\infty$. We can verify this convergence in probability from a mean-square convergence, by calculating the variance

$$\overline{\sigma}_\mu^2 = \mathbb{E}_\mu\big(\|\Phi_d(x) - \mathbb{E}_\mu(\Phi_d(x))\|^2\big).$$

The Markov inequality implies that if $\lim_{d\to\infty} \overline{\sigma}_\mu(\Phi_d(x))/\epsilon_d = 0$ then

$$\lim_{d\to\infty} \mathrm{Prob}\big(\|\Phi_d(X) - \mathbb{E}_\mu(\Phi_d(x))\| \leq \epsilon_d\big) = 0\,.$$

This means that when $d$ increases there is a probability converging to 1 that a realization of $X$ belongs to a microcanonical set computed from a single realization $\overline{x}$ with $y = \Phi_d(\overline{x})$:

$$\Omega_{d,\epsilon_d} = \big\{x \in \mathbb{R}^{\Lambda_d} : \|\Phi_d(x) - \Phi_d(\overline{x})\| \leq \epsilon_d\big\}\,.$$

In numerical calculations, we stop the gradient descents after a fixed number $n$ of iterations so that the resulting gradient descent measure is supported in a microcanonical set $\Omega_{d,\epsilon}$ for $\epsilon$ small enough. If $\epsilon/\overline{\sigma}_\mu^2(\Phi_d(x)) \gg 1$ then nearly all realizations of $X$ are included in $\Omega_{d,\epsilon}$. However, the microcanonical set may become too large and hence include points which are not typical realizations of $X$. We thus typically wait to reach a smaller $\epsilon$ width

Since $\Phi(x)$ is in a space of dimension $K$, Corollary 3.2 proves that reducing $\epsilon$ by a factor $\gamma$ reduces the maximum entropy of the microcanonical model by a factor of the order of $K \log \gamma$. In the extensive case, this maximum entropy is proportional to $d$ so the entropy reduction is negligible if $K\,|\log(\epsilon/\overline{\sigma}_\mu(\Phi_d(x))| \ll d$. In all numerical calculations of this paper $\epsilon/\overline{\sigma}_\mu(\Phi_d(x))$ is of the order of $10^{-3}$. We evaluate the concentration of $\Phi_d(X)$ by computing the normalized variance

$$\sigma^2_\mu(\Phi_d) = \frac{\mathbb{E}_\mu\big(\|\Phi_d(x) - \mathbb{E}_\mu(\Phi_d(x))\|^2\big)}{\mathbb{E}_\mu\big(\|\Phi_d(x)\|^2\big)}. \tag{48}$$

**Microcanonical gradient descent entropy.** Since $I = \mathbb{R}$, the gradient descent is initialized with a Gaussian white noise measure $\mu_0$ of variance $\sigma_0^2 = d^{-1}\|\overline{x}\|_2^2$. The convergence of the gradient descent algorithm to the microcanonical set is checked by verifying that for almost all Gaussian white realization $x_0$, after a sufficient large number $n$ of gradient steps

$$\|\Phi_d(x_n) - \Phi_d(\overline{x})\| \le \epsilon,$$

and hence $x_n \in \Omega_{d,\epsilon}$. Convergence issues may be due to existence of local minima or because the Hessian of $\Phi_d(x)$ is too ill-conditioned. Let $\mu_n$ be the resulting microcanonical gradient descent measure. If $\mu_n$ is supported in $\Omega_{d,\epsilon}$ then it has a smaller entropy than the maximum entropy microcanonical measure, which is uniform in $\Omega_{d,\epsilon}$. Theorem 3.7 gives an upper bound on the reduction of entropy.

**Model error.** Suppose that the restriction of $X$ to $\Lambda_a$ has a maximum entropy measure $\mu$ associated to a known energy $\Phi^\mu_d(x)$. This will be the case for Gaussian or Ising processes. The typical sets where the realizations of $X$ are almost all concentrated are sets where $\|\Phi^\mu_d(x) - \mathbb{E}_\mu(\Phi^\mu_d(x))\|$ is sufficiently small. In this case we can verify that the gradient descent microcanonical measure $\mu_n$ computed with a model energy $\Phi_d$ is also included in such a typical set with high probability. This concentration property is satisfied if the mean-square variation of the process energy $\mathbb{E}_{\mu_n}(\|\Phi^\mu_d(x) - \mathbb{E}_\mu(\Phi^\mu_d(x))\|^2)$ converges to 0 when $d$ increases. This convergence is evaluated by computing the concentration of $\Phi^\mu_d(x)$ around $\mathbb{E}_\mu(\Phi^\mu_d(x))$ for $\mu_n$:

$$e^2_{\mu_n}(\Phi_d) = \frac{\mathbb{E}_{\mu_n}\big(\|\Phi^\mu_d(x) - \mathbb{E}_\mu(\Phi^\mu_d(x))\|^2\big)}{\mathbb{E}_{\mu_n}\big(\|\Phi^\mu_d(x)\|^2\big)}. \tag{49}$$

If $\mu_n = \mu$ then $e^2_{\mu_n}(\Phi_d) = \sigma^2_\mu(\Phi^\mu_d)$ but the reverse is not true. It would be true only if the microcanonical gradient descent measure had a maximum entropy, which is not valid in general. On the other hand, if $e^2_{\mu_n}(\Phi_d) \gg \sigma^2_\mu(\Phi^\mu_d)$ then it indicates that there is a model error.

**5.2. Approximation of Gaussian processes.** We study approximations of stationary Gaussian random processes with gradient descent microcanonical models, defined with wavelet and scattering energy vectors.

We consider a scalar quadratic potential $Ux = |x \star h(u)|^2$ for $u \in \mathbb{Z}^2$. As in (18), we define a periodic filter $h_d(n) = \sum_{m \in \mathbb{Z}^2} h(n - md^{1/2})$ over square images of $d$ pixels and an energy

$$\Phi_d^\mu(x) = d^{-1} \|x \star h_d\|_2^2 = d^{-1} \sum_\omega |\widehat{x}(\omega)|^2 \, |\widehat{h}_d(\omega)|^2. \tag{50}$$

If $\inf_\omega |\widehat{h}(\omega)| > 0$ then we saw in (20) that microcanonical and macrocanonical models converge to a Gaussian stationary process $\mu$ over $\mathbb{Z}^2$ whose power spectrum is

$$P_\mu(\omega) = \beta^{-1} \, |\widehat{h}(\omega)|^{-2} \,. \tag{51}$$

In numerical experiments, we choose a discrete filter $h(n) = c\, e^{-|n|/\xi}$ with $\xi = 0.5$, whose Fourier transform satisfies for $\omega \in [-\pi, \pi]^2$

$$|\widehat{h}(\omega)|^2 = c^2 \sum_{m \in \mathbb{Z}^2} \big(\xi^2 + |\omega + 2m\pi|^2\big)^{-2}. \tag{52}$$

Figure 1(a) shows realizations of the Gaussian process of power spectrum $P_\mu(\omega)$, which is nearly the same as the maximum entropy microcanonical process computed with the scalar energy $\Phi_d^\mu$. Since $\Phi_d^\mu$ is an $\mathbf{l}^2$ energy, Theorem 3.9 proves that the gradient descent is not trapped in a local minima and thus converges to a microcanonical set of $\Phi_d^\mu$. This is verified by Table 1 where $e^2_{\mu_n}(\Phi_d^\mu) = \sigma^2_\mu(\Phi_d^\mu)$. However Figure 1(b) shows that realizations of the microcanonical gradient descent process are different from realizations of the original Gaussian process and hence of the maximum entropy microcanonical process. Figure 2(a,b) show that the maximum entropy microcanonical process has a power spectrum which is different from the spectrum of the microcanonical gradient descent process.

Observe that the power spectrum in Figure 2(a,b) are invariant by rotations in the Fourier plane. These rotations are orthogonal operators and they preserve the stationary mean which corresponds to the Fourier transform value at $\omega = 0$. If $\widehat{h}_d(\omega)$ is invariant by a rotation of $\omega$ then (50) implies that $\Phi_d^\mu(x)$ is invariant to these rotations, and Theorem 3.4 proves that $\mu^{\min}_{d,\epsilon}$ and $\mu_n$ are invariant to these rotations. This rotation invariance is not strictly valid at the highest frequencies because of the square grid sampling.

| | $\Phi_d = \Phi_d^\mu$ | $\Phi_d =$Wavelet $\mathbf{l}^2$ | $\Phi_d =$ Wavelet $\mathbf{l}^1$ | $\Phi_d =$ Scattering |
|---|---|---|---|---|
| $\dim(\Phi_d)$ | 1 | 40 | 40 | 114 |
| $\sigma^2_\mu(\Phi_d)$ | 5e-4 | 4e-3 | 4e-3 | 5e-3 |
| $e^2_{\mu_n}(\Phi_d)$ | 5e-4 | 2e-2 | 0.15 | 2e-2 |

Table 1. The first line gives the dimension of each energy vectors $\Phi_d(x)$. The next lines give the normalized variance $\sigma^2_\mu(\Phi_d)$ and the process energy concentration $e^2_{\mu_n}(\Phi_d)$, depending upon the microcanonical energy vector $\Phi_d$, for the Gaussian process (51).

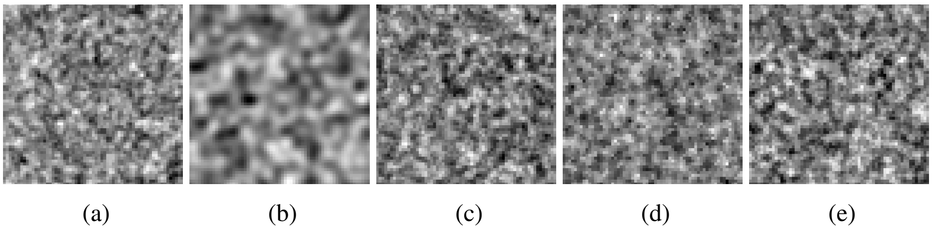



Figure 1. (a) Realization of the Gaussian process (51). (b) Realization of the microcanonical gradient descent computed with $\Phi_d(x) = \Phi_d^{\mu}(x) = \|x \star h\|_2^2$. (c) Realization computed with a vector $\Phi_d(x)$ of $\mathbf{l}^2$ wavelet norms. (d) $\Phi_d(x)$ is composed of $\mathbf{l}^1$ wavelet norms. (e) $\Phi_d(x)$ is a scattering transform.

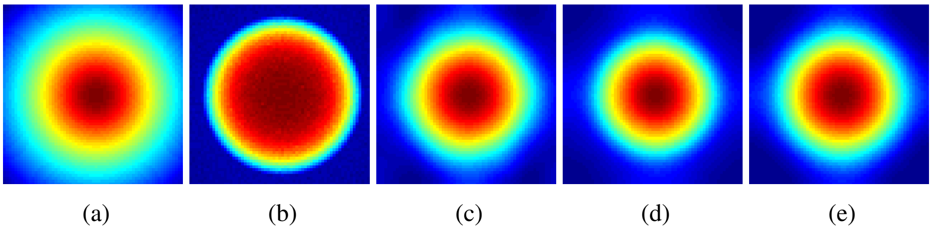



Figure 2. (a) Power spectrum of the original Gaussian process. (b) Estimation of the spectrum of a microcanonical gradient descent computed with the energy vector $\Phi_d(x) = \phi_d(x) = \|x \star h\|_2^2$. (c) The energy vector $\Phi_d(x)$ consists of $\mathbf{l}^2$ wavelet norms. (d) $\Phi_d(x)$ includes $\mathbf{l}^1$ wavelet norms. (e) $\Phi_d(x)$ includes $\mathbf{l}^1$ scattering norms.

**Wavelet $\mathbf{l}^2$ norms.** Let us now compute the gradient descent microcanonical measure $\mu_n$ with a wavelet $\mathbf{l}^2$ norm energy vector $\Phi_d$ in (42). We shall see that it can provide good approximations of Gaussian processes. The normalized variance $\sigma^2(\Phi_d)$ in Table 1 remains small which indicates that this energy vector remains concentrated around its mean. Figure 1(c) shows a realization of the resulting microcanonical gradient descent model and Figure 2(c) gives an estimation of the power spectrum of this stationary process. This power spectrum is now much closer to the original power spectrum.

To understand this, observe that wavelet $\mathbf{l}^2$ norms specify the signal energy in the different frequency bands covered by each band-pass wavelet filter $\widehat{\psi}_{j,q}(\omega)$:

$$\|x \star \psi_{j,q}\|^2 = \sum_{\omega} |\widehat{x}(\omega)|^2 \, |\widehat{\psi}_{j,q}(\omega)|^2. \tag{53}$$

The fact that the power spectrum remains nearly constant over the support of each $\widehat{\psi}_{j,q}$ is a consequence of Theorem 3.4(iii). Indeed, suppose that $Lx$ is a linear operator which performs a permutation of the values of $\widehat{x}(\omega_1)$ and $\widehat{x}(\omega_2)$, for two non-zero frequencies $\omega_1$ and $\omega_2$ such that $\widehat{\psi}_{j,q}(\omega_1) = \widehat{\psi}_{j,q}(\omega_2)$ for all $j, q$. It is an orthogonal operator which preserves the mean (zero frequency) and it is a symmetry of $\Phi_d$. Theorem 3.4(iii) implies that the gradient descent measure $\mu_n$ is also invariant to the action of $L$ and is thus a stationary process whose power spectrum is the same at $\omega_1$ and $\omega_2$. This property is approximately valid for any frequencies $\omega_1$ and $\omega_2$ located near the center of the support of each $\widehat{\psi}_{j,q}$, where it remains nearly constant and where all other $\widehat{\psi}_{j',q'}$ nearly vanish. It implies that the spectrum of $\mu_n$ remains nearly constant in these frequency domain.

The energy concentration $e^2_{\mu_n}$ in Table 1 is small although not as small as $\sigma^2_\mu(\Phi^\mu_d)$ which indicates the presence of a bias. To reduce this bias we must reduce the support size of each wavelet $\widehat{\psi}_{j,q}$ where the spectrum must remain nearly constant. Appropriate wavelet design can yield arbitrarily small errors when $d$ increases.

Besides having an appropriate power spectrum, these microcanonical gradient descent models are also nearly Gaussian processes. This can be shown with a phase symmetry argument, which is explained without a formal proof. The wavelet norms in (53) and hence $\Phi_d(x)$ are invariant if we preserve $|\widehat{x}(\omega)|$ but change the complex phase of $\widehat{x}(\omega)$ for $\omega \neq 0$. Arbitrary rotations of the Fourier complex phases which transform real signals into real signals are linear orthogonal operators which preserve the stationary mean. As a result, Theorem 3.4 proves that the gradient descent process is invariant to any such Fourier phase rotation. This means that Fourier transforms of realizations of these microcanonical gradient descent processes have phases which are independent and uniformly distributed. Given a fixed power spectrum, a standard result based on the central limit theorem proves that stationary random processes with independent and uniformly distributed Fourier phases converge to a Gaussian processes when the dimension $d$ goes to $\infty$ [21]. Under appropriate hypotheses, microcanonical gradient descent processes conditioned by $\mathbf{l}^2$ wavelet norms will thus converge to Gaussian processes.

**Wavelet $\mathbf{l}^1$ norms.** Maximum entropy models conditioned by wavelet $\mathbf{l}^1$ norms capture sparsity with Laplacian distributions but do not approximate Gaussian processes accurately. Figure 1(d) shows samples of the microcanonical gradient processes computed with a wavelet $\mathbf{l}^1$ norm energy (44). The $\mathbf{l}^1$ norm constraints produce wavelet coefficients which are more sparse than a true Gaussian process. It creates images which are more piece-wise regular than in Figure 1(c). Errors are also visible in the resulting power spectrum shown in Figure 2(d). Table 1 shows that the resulting model error $e^2_{\mu_n}$ for the $\mathbf{l}^1$ norm wavelet vector is about 10 times larger than with the $\mathbf{l}^2$ wavelet energy vector.

**Scattering energy.** The scattering energy vector (46) includes high order multiscale terms which can nearly reproduce the $\mathbf{l}^2$ norms of wavelet coefficients, as proved by Proposition 4.1. Table 1 gives the normalized variance $\sigma^2_\mu(\Phi_d(x))$ which shows that it concentrates nearly as well as wavelet $\mathbf{l}^2$ norm energy vectors, despite the fact that it is much larger. Figure 1(e) shows a realization of the scattering microcanonical gradient descent model and Figure 2(e) gives its power spectrum. It is nearly as precise as the $\mathbf{l}^2$ norm microcanonical model and the model error $e^2_{\mu_n}$ in Table 1 has about the same amplitude.

**5.3. Ising processes.** We consider a two-dimensional Ising process with no outside magnetization, over a two-dimensional square lattice with periodic boundary conditions. We denote by $x(u)$ the spin values in $\{-1, 1\}$. The Ising probability of a configuration $x$ is

$$p(x) = \mathcal{Z}^{-1} \exp\left(-\beta\, \phi_d(x)\right) \text{ with } \phi_d(x) = d^{-1} \sum_{u \in \Lambda_d} \sum_{u' \in \mathcal{N}_u} x(u)x(u'), \tag{54}$$

where $\mathcal{N}_u$ is the 4 point neighborhood of $x(u)$ in the two-dimensional grid. The constant $\beta = (k_B T)^{-1}$ is the inverse temperature scaled by the Boltzmann constant $k_B$. In two dimension, the free energy can be exactly computed with the method of Onsager [37]. It has a phase transition when $T$ reaches a critical value $T_c \approx 2.27$. We study the approximation of Ising for several values of the temperature.

The complex behavior of Ising arises from the conjunction of the quadratic Hamiltonian with the binary constraint. This binary condition may be replaced by a condition on a fourth order moment to obtain the same critical behavior but we shall impose it here through first and second order moments. For all $x \in \mathbb{R}^d$, one has $\|x\|_2 \le \|x\|_1 \le \sqrt{d}\,\|x\|_2$, and $\|x\|_1 = \sqrt{d}\,\|x\|_2$ if and only if $|x(u)|$ is constant. It follows that

$$\forall u\ ,\ x(u) = \pm 1 \Leftrightarrow \|x\|_1 = \|x\|_2^2 = d\ .$$

We can thus impose that $x$ is binary by adding $d^{-1}\|x\|_2^2$ and $d^{-1}\|x\|_1$ into the energy vector. The resulting microcanonical interaction energy for $x \in \mathbb{R}^{\Lambda}_d$ is

$$\Phi^{\mu}_d(x) = \left\{d^{-1}\|x\|_2^2\, ,\, d^{-1}\|x\|_1\, ,\, \phi_d(x)\right\}. \tag{55}$$

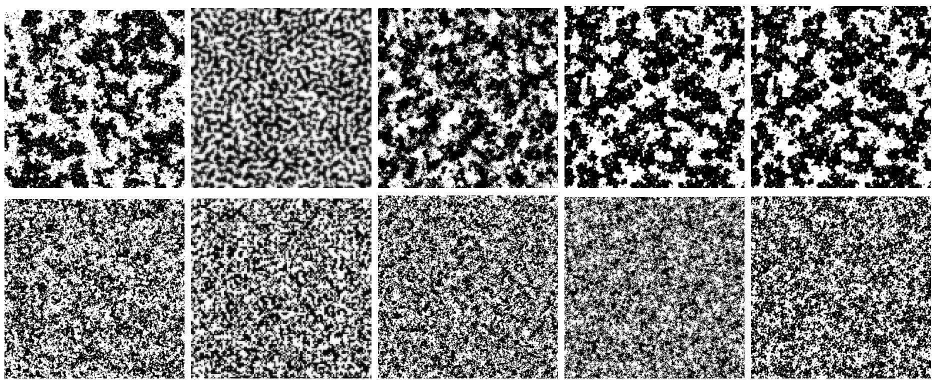
If we remove the $\mathbf{l}^1$ term, this energy is quadratic and the maximum entropy model is therefore a stationary Gaussian process.

The Ising model has a phase transition at the critical temperature $T_c \approx 2.27$, from an "ordered" to a "disordered" state. The spin spatial correlation exhibits a characteristic scale $\xi(T)$ for $T > T_c$ and $\mathbb{E}\{X(u)X(u+r)\} \simeq e^{-|r|/\xi(T)}$ [29], with $\xi(T_c) = 0$. The correlation is self-similar at $T = T_c$ and $\mathbb{E}\{X(u)X(u+r)\} \simeq |r|^{-1/2}$.

Figure 2(a) gives two realizations of Ising for a large temperature (bottom) and a temperature just above the critical temperature (top). Figure 2(b) shows realizations of the microcanonical gradient descent process computed with the Ising energy vector $\Phi^{\mu}_d$. The first column of Table 2 shows that $e^2_{\mu_n}(\Phi^{\mu}_d) \gg \sigma^2(\mu(\Phi^{\mu}_d)$ which means that the microcanonical gradient descent does not converge to a microcanonical set for $\epsilon$ small. Near the critical temperature, the gradient descent microcanonical model is unable to recover low-frequency long-range structures which appear in Ising. This is due to a well-known instability near criticality.

| | $\Phi_d = \Phi_d^\mu$ | $\Phi_d$ = Wavelet $\mathbf{l}^2$ | $\Phi^d$ = Wavelet $\mathbf{l}^1$ | $\Phi_d$ = Scattering |
|---|---|---|---|---|
| $\dim(\Phi_d(x))$ | 3 | 42 | 42 | 116 |
| $\sigma_\mu^2(\Phi_d)$, $T = 2.2$ | 6e-6 | 3e-4 | 4e-4 | 6e-4 |
| $e_{\mu_n}^2(\Phi_d)$, $T = 2.2$ | 2e-2 | 7e-2 | 5e-2 | 9e-3 |
| $\sigma_\mu^2(\Phi_d)$, $T = 3$ | 3e-6 | 2e-5 | 4e-5 | 4e-5 |
| $e_{\mu_n}^2(\Phi_d)$, $T = 3$ | 7e-3 | 4e-2 | 5e-2 | 5e-3 |

Table 2. The first line gives the dimension of each energy vectors $\Phi_d(x)$. We consider two Ising processes (54), computed near the critical temperature $T = 2.2$ and at a larger temperature $T = 3$. The table gives the normalized variance $\sigma_\mu^2(\Phi_d)$ and the Ising energy concentration $e_{\mu_n}^2(\Phi_d)$, for different $\Phi_d(x)$.



(a) (b) (c) (d) (e)

Figure 3. (a) Realizations of an Ising process near the critical temperature $T = 2.2$ (top), and for $T = 3$ (bottom). (b) Realizations computed with the microcanonical gradient descent with $\Phi_d = \Phi_d^\mu$. (c) $\Phi_d(x)$ includes $\mathbf{l}^2$ wavelet norms. (d) $\Phi_d(x)$ includes $\mathbf{l}^1$ wavelet norms. (e) $\Phi_d(x)$ includes $\mathbf{l}^1$ scattering norms.

**Renormalization and wavelets.** As in Wilson renormalization group, wavelets separate the frequency components of $x$ into dyadic frequency annulus. Relations between wavelets and renormalization group decompositions were studied by Battle [5]. In the following, we give a qualitative argument to explain how to approximate the Ising potential with wavelet norms.

Since $x(u) \in \{-1, 1\}$, for an integer $p$

$$x(u)x(u') = 1 - 2^{-1}\,|x(u) - x(u')|^p$$

so we can rewrite the Ising energy $\phi_d(x) = d^{-1} \sum_{u \in \Lambda_d} \sum_{u' \in \mathcal{N}_u} x(u)x(u')$ satisfies

$$d - \phi_d(x) = 2^{-1} \sum_{u \in \Lambda_d} \sum_{u' \in \mathcal{N}_u} |x(u) - x(u')|^p = \|\Delta_1 x\|_p^p + \|\Delta_2 x\|_p^p, \tag{56}$$

with $\Delta_1 x(u_1, u_2) = x(u_1, u_2) - x(u_1, u_2 - 1)$ and $\Delta_2 x(u_1, u_2) = x(u_1, u_2) - x(u_1 - 1, u_2)$.

The equivalence of $\mathbf{l}^p$ norms of increments and $\mathbf{l}^p$ norms of wavelet coefficients is established in [36]. For any $p > 1$ there exists $A_p > 0$ and $B_p > 0$ so that for any $x \in \mathbf{l}^2(\mathbb{Z}^2)$

$$A_p \sum_{j,q} 2^{-jp} \|x \star \psi_{j,q}\|_p^p \le \|\Delta_1 x\|_p^p + \|\Delta_2 x\|_p^p \le B_p \sum_{j,q} 2^{-jp} \|x \star \psi_{j,q}\|_p^p. \tag{57}$$

For $p = 1$ the upper-bound remains valid but to get a lower-bound we must replace the sum over $j, q$ by a sup operator. However, we conjecture that there exists $A_1$ which verifies the lower bound for $p = 1$ when the values of $x(u)$ are restricted to $\{-1, 1\}$. With equations (56) and (57) one can approximate the Ising energy $\phi_d(x)$ with discrete wavelet $\mathbf{l}^p$ norms computed at all scales $2^j \le 2^J = d$. We limit the maximum scale $2^J$ independently of $d$, which is set to be the largest correlation length of the process.

As in Section 5.3, we capture the fact that $x(u) \in \{-1, 1\}$ by including a condition on $d^{-1}\|x\|_1$ and $d^{-1}\|x\|_2^2$. The resulting energy vector for $p = 1$ and $p = 2$ is

$$\Phi_d(x) = \left\{ d^{-1}\|x\|_2^2 \,, d^{-1}\|x\|_1 \,,\, d^{-1}\,\|x \star \psi_{j,q}\|_p^p \right\}_{j \le J, q \le Q}. \tag{58}$$

Table 2 shows the normalized variance $\sigma^2(\Phi_d)$ is smaller at high temperature than near critical temperature but the separation of scale still provides a high concentration of $\Phi_d(x)$ for an Ising process, close to the critical temperature. Figure 3(c,d) show realizations of a microcanonical gradient descent Ising model computed with the wavelet energy (58) for $p = 1$ and $p = 2$. Near critical temperature, the microcanonical gradient descent still converges where as it was not the case when the energy was calculated directly with the Ising Hamiltonian energy $\phi_d(x)$ in Figure 3(b). The scale separation avoids having an ill-conditioned gradient descent. The Ising approximation with an $\mathbf{l}^2$ energy vector for $p = 2$ amounts to compute a Gaussian approximation of Ising, which is not precise, when we are close to the critical temperature [28]. One can indeed visualize important differences with the statistical distribution of original Ising in Figure 3(a). Table 2 shows that the model error $e^2_{\mu_n}$ is smaller at higher temperature.

The Ising approximation with an $\mathbf{l}^1$ energy vector has about the same error as the model computed with an $\mathbf{l}^2$ energy vector. Near the critical temperature, the microcanonical models obtained with $\mathbf{l}^1$ wavelets norms shown in Figure 3(d) are more piecewise regular than the ones in Figure 3(c) obtained with wavelet $\mathbf{l}^2$ norms. This is due to the wavelet coefficient sparsity imposed by these $\mathbf{l}^1$ norms.

**Scattering energy.** A scattering energy vector is defined for Ising process, by complementing the scattering energy vector (46) with $\mathbf{l}^1$ and $\mathbf{l}^2$ norms of $x$ in order to impose that $x(u)$ takes binary values:

$$\Phi_d(x) = \Big\{d^{-1}\|x\|_2^2\,,\ d^{-1}\|x\|_1\,,\ d^{-1}\sum_{u\in\Lambda_d} x(u)\,,$$
$$d^{-1}\,\|x \star \psi_{j,q}\|_1\,,\ d^{-1}\,\||x \star \psi_{j,q}| \star \psi_{j',q'}\|_1\Big\}_{j,j'\leq J,q,q'\leq Q}. \quad (59)$$

Table 2 shows that the normalized variance of the scattering energy is about twice larger than for $\mathbf{l}^2$ wavelet energy vectors. Figure 3(e) shows realizations of microcanonical gradient descent models computed with this scattering energy vector. They are visually difficult to distinguish from realization of the original Ising process above the critical temperature and close to the critical temperature. Table 2 shows that the model error $e^2_{\mu_n}$ is about 10 times smaller than with $\mathbf{l}^2$ or $\mathbf{l}^1$ wavelet energies.

These numerical experiment seem to indicate that scattering microcanonical gradient descents can provide accurate model of Ising even close to critical temperature. However, this needs to be sustained by a better mathematical of these approximations, by analyzing the preservation of symmetries.

**5.4. Point processes.** Point processes provide powerful models of stochastic geometry, with applications in many areas of astrophysics, neuroscience, finance and computer vision. Realizations of point processes have a support reduced to isolated points. We first show that this sparsity can be captured by wavelet $\mathbf{l}^1$ norms. We then study approximations of point processes and shot noises with microcanonical models defined by scattering coefficients.

**Support from wavelet $\mathbf{l}^1$ norms.** We prove that wavelet $\mathbf{l}^1$ norms capture important geometric properties of the support of point processes. Young's inequality implies that

$$\|x \star \psi_{j,q}\|_1 \leq \|x\|_1\,\|\psi_{j,q}\|_1.$$

If $x$ is a Dirac in $\Lambda_d$ then this inequality is an equality. Conversely, the following theorem, proved in Appendix I proves that if this inequality is an equality then $x$ is a sum of Diracs, with conditions on their distances. The inner product and norm of $v$ and $v'$ in $\mathbb{R}^\ell$ is written $v.v'$ and $\|v\|$.

We suppose that wavelets are defined from a mother wavelet $\psi(u)$ which is continuous with $\psi(0) \neq 0$. We suppose that $\psi(u) = |\psi(u)|\, e^{i\,\varphi(\xi.u)}$ where $\xi \in \mathbb{R}^\ell$ and the complex phase $\varphi$ is a bi-Lipschitz function. We may choose linear phase $\varphi(\xi.u) = \xi.u$. This wavelet is rotated and dilated $\psi_{j,q}(u) = 2^{-j\ell}\psi(2^{-j} r_q^{-1} u)$, where the $r_q$ are $Q \geq \ell$ different rotations in $\mathbb{R}^\ell$. The following theorem applies to these wavelets.

**Theorem 5.1.** (i) *If* $\|x \star \psi_{j,q}\|_1 = \|x\|_1\, \|\psi_{j,q}\|_1$ *then* $x$ *is non-zero at* $u$ *and* $u'$ *only if* $\xi_q.(u-u') = 0$ *with* $\xi_q = r_q \xi$ *or if* $|\xi_q.(u-u')| \geq C\, 2^j$ *, where* $C > 0$ *does not depend on* $x$.

(ii) *Suppose that* $\psi$ *has a compact support, and that* $x$ *has a support which is a union of isolated points with distances larger than* $\Delta$. *If* $x'$ *satisfies*

$$\forall q \leq Q\,,\ \forall j \leq \log_2 \Delta\,,\ \|x' \star \psi_{j,q}\|_1 = \|x \star \psi_{j,q}\|_1 \text{ and } \|x'\|_1 = \|x\|_1 \quad (60)$$

*then the support of* $x'$ *is a set of isolated points of distances larger than* $C\,\Delta$, *where* $C > 0$ *does not depend on* $x$.

In dimension $\ell = 2$, property (i) of Theorem 5.1 proves that the support of $x$ is included in straight lines perpendicular to $\xi_q$, whose distances are larger than $C\, 2^j$. If this is valid for several $q$ then the support is included over intersections of non-parallel lines and hence reduced to isolated points, as proved by property (ii).

If $x$ is a realization of a point process, its support is a union of isolated points whose minimum distance depends the point process distribution. If we construct an $\epsilon = 0$ microcanonical model with wavelet $\mathbf{l}^1$ norms then property (ii) proves that all realizations of this microcanonical model will also be a point process with a similar separation between points.

**Microcanonical models of point processes.** We study microcanonical models of point processes with wavelet $\mathbf{l}^1$ norms and scattering coefficients. A point process $N$ on $\mathbb{R}^\ell$ is a measure whose support is composed of isolated points. Second-order point processes [8] are those satisfying $\mathbb{E}[N(C)^2] < \infty$ for all bounded Borel sets $C \subset \mathbb{R}^\ell$. If $N$ is a stationary, second-order point process then one can define its associated Bartlett spectral measure [8] $P_N$, which generalizes the power spectrum of second-order stationary processes.

Given a non-negative stationary process $\lambda(t)$, $t \in \mathbb{R}^\ell$, a Cox process $N$ is defined as a Poisson process conditional on $\lambda$ with intensity $\lambda(t)$. Important geometric information of $N$ is captured by its Bartlett power spectrum, which satisfies $P_N(d\omega) = P_\lambda(d\omega) + \mathbb{E}(\lambda)\,\delta(d\omega)$ [8]. Shot noises are classes of random processes defined by convolutions of point processes with a filter $h(t)$

$$X(t) = N \star h(t)\,.$$

The filter $h(t)$ can be interpreted as a pattern which is randomly translated at point locations and added. It may also be the transfer function of a detector measuring the point-process. In this case, the power spectrum of $X$ is $P(d\omega) = P_N(d\omega)\,|\widehat{h}(\omega)|^2$, which mixes the geometric information of $N$ with the profile of the filter $h$. We will show that they can be disentangled by a wavelet scattering transform.

The loss of information in the power spectrum is due to the fact that it does not measure scale interactions. When there is a scale separation between $N$ and $h$, i.e.

$$\mathbb{E}(\lambda)^2 \gg \int u^2 |h(u)|^2 \, du \tag{61}$$

then for sufficiently small scales $2^j$, one can verify [11] that

$$|X \star \psi_{j,q}| = |N \star (\psi_{j,q} \star h)| \approx N \star |\psi_{j,q} \star h| \tag{62}$$

with high probability, due to the fact that the events in $N$ rarely interact at spatial scales $j$ such that $2^j \ll \mathbb{E}(\lambda)$. From this approximation, it follows that for sufficiently large scale gap $j' \gg j$, we have

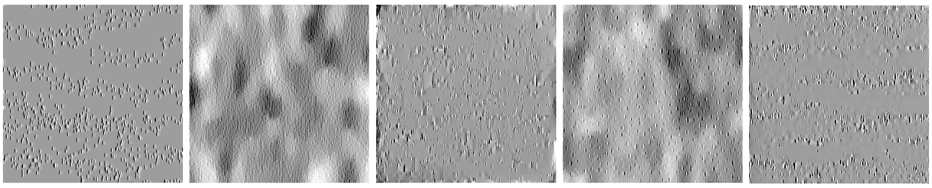
$$||X \star \psi_{j,q}| \star \psi_{j',q'}| \approx C_{j,q} |N \star \psi_{j',q'}| \,, \tag{63}$$

since $|\psi_{j,q} \star h| \star \psi_{j',q'} \approx C_{j,q} \delta \star \psi_{j',q'}$. Second order scattering coefficients, indexed with pairs $(j, q, j', q')$, thus provide measurements that convey spectral information about the point process $N$ as $(j', q')$ varies, disentangled from the spectral information of $h$.

We illustrate this phenomena by considering a two-dimensional Cox point process $N(u)$, whose rate $\lambda(u)$ is a stationary Gaussian process whose power spectrum is concentrated in the low-frequencies, and with an integral scale of 100 pixels. This Cox process is convolved with a pattern $h(u)$ with zero mean and small spatial support of 5 pixels. We build microcanonical models with energy vectors $\Phi_d(x)$ defined by wavelet $\mathbf{l}^1$ norms or scattering coefficients, computed up to a maximum scale $2^J$. For the shot noise measure $\mu$ shown in Figure 4(a), Table 3 gives the normalized variances $\sigma_\mu^2 = \mathbb{E}_\mu(\|\Phi_d(x) - \mathbb{E}(\Phi_d(x))\|^2)/\|\mathbb{E}_\mu(\Phi_d(x))\|^2$ as a function of the maximum scale $2^J$. Although the size of scattering vectors for large $J$ becomes relatively large, the normalized variance remains small which proves that these energy vectors remain concentrated around their mean, for images of size $d = 256^2$. We can thus define microcanonical models from an energy vector $\Phi_d(\overline{x})$ calculated from the realization $\overline{x}$ shown in Figure 4(a).

| | $J=2$ | | $J=4$ | | $J=6$ | |
|---|---|---|---|---|---|---|
| | $\sigma^2_\mu(\Phi_d)$ | $\dim(\Phi_d)$ | $\sigma^2_\mu(\Phi_d)$ | $\dim(\Phi_d)$ | $\sigma^2_\mu(\Phi_d)$ | $\dim(\Phi_d)$ |
| $\Phi_d$: Wavelet $\mathbf{l}^1$ | $4 10^{-6}$ | 21 | $3\,10^{-6}$ | 38 | $3\,10^{-6}$ | 52 |
| $\Phi_d$: Scattering | $8\,10^{-6}$ | 88 | $10^{-5}$ | 422 | $10^{-5}$ | 580 |

Table 3. Estimated normalized variance for wavelet $\mathbf{l}^1$ norm and scattering energy vectors $\Phi_d$, at different maximum scales $2^J$. They are computed for a shot noise of size $d = 256^2$ defined from a Cox point process. Figure 4(a) shows a realization.



(a) (b) (c) (d) (e)

Figure 4. (a) Realization of a shot noise computed with a Cox process. (b), (c) Realizations of a gradient descent process, computed with an energy $\Phi_d$ including wavelet $\mathbf{l}^1$ norms of maximum scale respectively $2^J = 8$ and $2^J = 64$. (d), (e) Same computed with an energy $\Phi_d$ including scattering $\mathbf{l}^1$ norms of maximum scale respectively $2^J = 8$ and $2^J = 64$.

Figure 4 gives realizations of microcanonical gradient descent models computed from wavelet $\mathbf{l}^1$ norms and scattering energies, at different maximum scales $2^J$. Figure 4(b,d) are computed with $2^J = 8$. These microcanonical models can only capture sparsity properties up to this maximum scale. At larger scale, the entropy maximisation creates Gaussian random process like variations having a uniform low-frequency spectrum. Figure 4(c,e) are microcanonical realizations computed at a larger maximum scale $2^J = 64$. In this case, wavelet $\mathbf{l}^1$ norm and scattering microcanonical models capture the point process sparsity. The geometry of the shot noise is defined by the stationary rate $\lambda(u)$ which has relatively high frequency oscillations vertically but low frequency variations horizontally. The scattering model Figure 4(e) captures this distribution thanks to second order coefficients. This is not the case for the $\mathbf{l}^1$ norm model in Figure 4(c) which can not reproduce the low-frequency horizontal alignments.

**5.5. Image and audio texture synthesis.** An image or an audio texture is usually modeled as the realization of a stationary process. Modeling textures amounts to compute an approximation of this stationary process given a single realization. A texture synthesis then consists in calculating new realizations from this stochastic model, which are hopefully perceptually identical to the original texture sample, although different if considered as deterministic signals. As opposed to the Gaussian, Ising or point process examples, since we do not know the original stochastic process, perceptual comparisons are the only criteria used to evaluate a texture synthesis algorithm. Microcanonical models can be considered as texture models computed from an energy function $\Phi_d(x)$ which concentrate close to its mean. We review previous work and give results obtained with a scattering microcanonical gradient descent model.

Geman and Geman [24] have introduced macrocanonical models based on Markov random fields. They provide good texture models as long as these textures are realizations of random processes having no long range correlations. Several approaches have then been introduced to incorporate long range correlations. Heeger and Bergen [26] capture texture statistics through the marginal distributions obtained by filtering images with oriented wavelets. This approach has been generalized by the macrocanonical Frame model of Mumford and Zhu [49], based on marginal distributions of filtered images. The filters are optimized by trying to minimize the maximum entropy conditioned by the marginal distributions. Although the Cramer-Wold theorem proves that enough marginal probability distributions characterize any random vector defined over $\mathbb{R}^d$ the number of such marginals is typically intractable, which limits this approach.

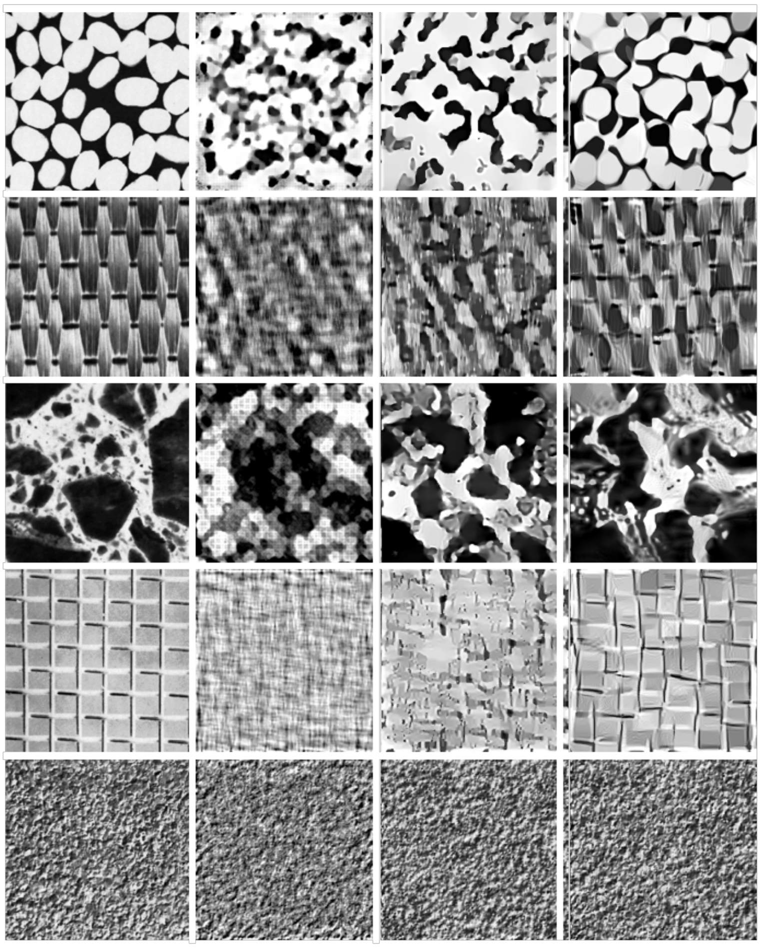
Portilla and Simoncelli [39] made important improvements to these texture models, with wavelet transforms. They capture the correlation of the modulus of wavelet coefficients with a covariance matrix which defines an energy vector $\Phi_d(x)$. Although they use a macrocanonical maximum entropy formalism, their algorithm computes a microcanonical estimation from a single realization, with alternate projections as opposed to a gradient descent. This approach was extended to audio textures by McDermott and Simoncelli [35]. A scattering representation is related to Portilla and Simoncelli model but covariance coefficients are replaced by a much smaller number of scattering $\mathbf{l}^1$ norms.

Excellent texture synthesis have recently been obtained with deep convolutional neural networks. In [23], the authors consider a deep VGG convolutional network, trained on a large-scale image classification task. The energy vector $\Phi_d(x)$ is defined as the spatial cross-correlation values of feature maps at every layer of the VGG networks. This energy vector is calculated on a particular texture image. Texture syntheses of very good perceptual quality are calculated with a gradient descent microcanonical algorithm initialized on random noise. However, the dimension of this energy vector $\Phi_d(x)$ is larger than the dimension $d$ of $x$. These estimators are therefore not statistically consistent and have no asymptotic limit.

In the following, we give results obtained with different wavelet microcanonical models computed on a collection of natural image and auditory textures. The Brodatz image texture dataset[2] consists of 155 texture classes, with a single $512 \times 512$ sample

[2]Available at `http://sipi.usc.edu/database/database.php?volume=textures`.

per class. Auditory textures are taken from McDermott and Simoncelli [35], which contains 1 second samples of different sounds.



(a) (b) (c) (d)

Figure 5. (a) Original texture. (b) texture synthesized with a microcanonical gradient descent model with a vector $\Phi_d(x)$ of wavelet $\mathbf{l}^2$ norms. (c) $\Phi_d(x)$ has wavelet $\mathbf{l}^1$ norms. (d) $\Phi_d(x)$ has wavelet scattering coefficients.

Since we have a single realization of each texture, we can not compute the concentration properties of energy vectors over these textures. Figure 5(a) gives input examples $\overline{x}$ corresponding to realizations of different stationary processes $X(u)$. Figure 5(b) shows texture samples obtained with a microcanonical gradient descent computed with an energy vector $\Phi_d(x)$ of wavelet $\mathbf{l}^2$ norm. It provides a good model for the bottom texture which is nearly Gaussian but it otherwise destroys the texture geometry. Figure 5(c) displays textures obtained with a vector $\Phi_d(x)$ of wavelet $\mathbf{l}^1$

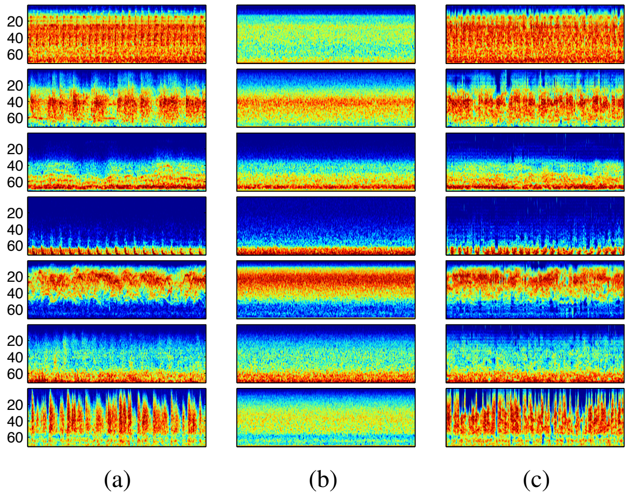



Figure 6. (a) Spectrograms of original audio textures produced (from to to bottom) by jack-hammer, applause, wind, helicopter, sparrows, train, rusting paper. (b) Spectrograms of an audio texture synthesized with a microcanonical gradient descent model with a vector $\Phi_d(x)$ of wavelet $\mathbf{l}^2$ norms. (c) Spectrogram produced with a vector $\Phi_d(x)$ of wavelet scattering coefficients.

norms. Their wavelet coefficients are more sparse than in Figure 5(b) which produces more "piecewise regular" images, but it does improve the texture geometry. On the contrary, scattering microcanonical textures in Figure 5(d) have a geometry which is much closer to original textures. Scattering coefficients can be interpreted as convolutional deep neural networks computed with predefined wavelet filters [10] as opposed to filters learned on a supervised image classification problem as in VGG.

The reconstruction of auditory textures is computed with a one-dimensional Gabor wavelet transform [9] with $Q = 12$ scales per octave. Auditory textures have a rich mixture of homogeneous and impulsive, transient components, as well as amplitude and frequency modulation phenomena. Figure 6(a) displays the spectrograms of original auditory textures $\bar{x}$. Figure 5(b) shows the spectrogram of Gaussian texture models calculated with a microcanonical gradient descent computed with an energy vector $\Phi_d(x)$ of wavelet $\mathbf{l}^2$ norm. The global spectral energy is preserved but the time variations which destroys ability to recognize these audio textures. On the contrary, Figure 5(c) shows that audio textures synthesized with a scattering energy vector have spectrograms with the same type of time intermittency as the original textures. The resulting audio textures are perceptually difficult to distinguish from the original ones.

Synthesis from scattering energy vectors can also destroy some certain structures which affect their perceptual quality. This is the case for speech or music backgrounds which have harmonic alignments which are not reproduced by scattering coefficients. Deep convolutional network reproduce image and audio textures of better perceptual quality than scattering coefficients, but use over 100 times more parameters. Much smaller models providing similar perceptual quality can be constructed with wavelet phase harmonics for audio signals [34] or images [48], which capture alignment of phases across scales. However, understanding how to construct low-dimensional multiscale energy vectors to approximate random processes remains mostly an open problem.

## 6. Conclusion

This paper shows that gradient descent microcanonical models computed with multiscale energy vectors can provide powerful models to approximate large classes of stationary processes. Realizations of such models are calculated with a gradient descent algorithm which is much faster than MCMC algorithms, used to sample from macrocanonical models.

We introduced a mathematical framework to analyze the statistical and algorithmic properties of these microcanonical gradient descent models. Our analysis reveals that, whereas mircocanonical gradient descent measures do not generally agree with the microcanonical maximum entropy measure, they have rich regularities through shared symmetries, and, under appropriate conditions, are shown to converge to the microcanonical ensemble. In the high-dimensional setting, gradient descent microcanonical models are therefore valid alternatives to classic macrocanonical and microcanonical maximum entropy measures, thanks to their computational tractability.

However, many mathematical questions remain open. For instance, on the convergence properties of this gradient descent algorithm, on the choice of the energy vector to obtain accurate approximations of random processes, and on the extension to locally stationary processes.

**Acknowledgements.** SM: This work was supported by the ERC grant InvariantClass 320959. JB: This work was partially supported by the Alfred P. Sloan Foundation, by NSF RI-1816753, and by Samsung DMC. We thank Zhengdao Chen and Ofer Zeitouni for valuable comments and fixes to the current manuscript, and the anonymous referees for their high-quality, valuable feedback.

## A. Proof of Theorem 3.1

**A.1. Proof of Part (i).** The main technical challenge to prove (26) is to show that assumption (C) is sufficient to guarantee that $|J\Phi_d x|^{-1}$ is integrable. Since $\Phi_d$ is

Lipschitz from assumption (A), the coarea formula proves that for any integrable function $g(x)$

$$\int_B g(x)\,|J\Phi_d x|\,dx = \int_{\mathbb{R}^K}\int_{\Phi_d^{-1}(y)} g(x)\,d\mathcal{H}^{d-K}(x)\,dy\,. \tag{64}$$

In order to apply (64) to $|J\Phi_d(x)|^{-1}$ and obtain the expression of $H(\mu^{\mathrm{mi}}_{d,\epsilon})$, we need to show that $|J\Phi_d(x)|^{-1}$ is integrable in $\Phi^{-1}_{d,\epsilon}(y)$. Using the notation for each Jacobian column (22), we verify that $|J\Phi_d(x)|$ satisfies

$$|J\Phi_d(x)| \geq d^{-\ell}\max\left\{|\det[JU(\bar{X}_1),\ldots,JU(\bar{X}_K)]|,\ldots\right.\\ \left.\ldots,|\det[JU(\bar{X}_{\tilde{d}+1}),\ldots,JU(\bar{X}_{\tilde{d}+K})]|\right\}, \tag{65}$$

where $\bar{X}_i$ is a projection of $x$ onto disjoint subsets of $2\Delta+1$ coordinates, and $\tilde{d} \geq d(2\Delta+1)^{-1} = \Theta(d)$.

We will show that for $d$ large enough and arbitrary $R>0$,

$$\int_{|x|_\infty<R} |J\Phi_d(x)|^{-1}dx < \infty\,, \tag{66}$$

by interpreting (66) as proportional to the expected value of $\mathbb{E}_{X\sim\mathrm{Unif}(d,R)}|J\Phi_d(X)|^{-1}$. Since $\Phi^{-1}_{d,\epsilon}(y)$ is a compact set thanks to assumption (B), it is bounded, so $\Phi^{-1}_{d,\epsilon}(y) \subseteq \{x\,;\,|x|_\infty < R\}$ for some $R$, which proves that $|J\Phi_d(x)|^{-1}$ is integrable in $\Phi^{-1}_{d,\epsilon}(y)$.

For that purpose, let us prove that assumption (C) from (25) is sufficient to guarantee (66). If $F_V(y)$ denotes the cumulative distribution function of a random variable $V$, and $Y$ denotes the r.v. $Y = |\det[JU(\bar{X}_1),\ldots,JU(\bar{X}_K)]|$, we first observe that thanks to (65) it is sufficient to show that

$$F_Y(y) \lesssim y^\eta\ ,\ \text{for some } \eta>0\ ,\ (y\to 0)\,. \tag{67}$$

Indeed, since $V = |J\Phi_d(X)| \geq \max(Y_1,\ldots,Y_{\tilde{d}})$ with $Y_i$ independent and identically distributed, we have that

$$F_V(y) \leq F_Y(y)^{\tilde{d}} \simeq y^{\eta\tilde{d}}\,.$$

It follows that

$$\mathbb{E}_{X\sim\mathrm{Unif}(d,R)}|J\Phi_d(X)|^{-1} \leq \int v^{-1}f_V(v)\,dv = C + \int_0^{R'} v^{-2}F_V(v)\,dv < \infty$$

as soon as $\tilde{d}\eta > 1$, which will happen for large enough $d$.

Let us thus prove (67) by induction on $K$. When $K=1$, $V = |\det JU(\bar{X}_1)| = |JU(\bar{X}_1)|$ and assumption (C) directly implies that

$$F_V(y) = P(V\leq y) \lesssim y^\eta\,.$$

Now, suppose (67) is true for $K-1$ and let us prove it for $K$. We use the following lemma:

**Lemma A.1.** *We say that a bounded random vector $Z$ in $B(K,R) \subset \mathbb{R}^K$ has property $(*)$ if there exists $\eta > 0$ such that*

$$\forall\ \mathcal{S} \subset \mathbb{R}^K \text{ Lebesgue measurable}\ ,\ P(Z \in \mathcal{S}) \lesssim |\mathcal{S}|^\eta\ .$$

*If $Z$ has property $(*)$ and $K > 1$, then $Z_H$, the orthogonal projection of $Z$ onto any hyperplane, also has property $(*)$, and*

$$\mathbb{E}\big(\|Z\|^{-\eta}\big) < C_{R,\eta}\ . \tag{68}$$

Before proving the lemma, let us conclude with (67). By denoting $Z_i = JU(\overline{X}_i)$, $i = 1,\ldots,K$, and assuming $\|Z_1\| > 0$, one Gram–Schmidt iteration yields

$$|\det\big[Z_1,\ldots,Z_K\big]| = \|Z_1\||\det\big[\widetilde{Z}_2,\ldots,\widetilde{Z}_K\big]|\ ,$$

where $\widetilde{Z}_i$ is the projection of $Z_i$ onto the orthogonal complement of $Z_1$. Using assumption (C), we use Lemma A.1 to observe that $\widetilde{Z}_i$, $i = 2,\ldots,K$ also satisfies assumption (C), since we compute it with an orthogonal projection that depends only on $Z_1$, which is independent from all the $Z_i$, $i \geq 2$. Thus by induction hypothesis and using (68) we obtain

$$\begin{aligned}
F_Y(y) &= P\big(|\det\big[Z_1,\ldots,Z_K\big]| \leq y\big) \\
&= P\big(\|Z_1\||\det\big[\widetilde{Z}_2,\ldots,\widetilde{Z}_K\big]| \leq y\big) \\
&= \mathbb{E}_{Z_1} P\big(|\det[\widetilde{Z}_2,\ldots,\widetilde{Z}_K]| \leq y\|Z_1\|^{-1} \mid Z_1\big) \\
&\leq \mathbb{E}_{Z_1} y^\eta \|Z_1\|^{-\eta} \lesssim y^\eta\ ,
\end{aligned}$$

which proves (67).

Let us finally prove Lemma A.1. Let $\mathcal{S}_H$ be a measurable set in a given hyperplane $H$ of dimension $K-1$, and let $\widetilde{\mathcal{S}} = \mathcal{S}_H \times (-R,R)$ be the corresponding cylinder in $B(K,R)$. By definition, we have

$$P(Z_H \in \mathcal{S}_H) = P(Z \in \widetilde{\mathcal{S}}) \leq |\widetilde{\mathcal{S}}|^\eta = |\mathcal{S}_H|^\eta (2R)^\eta$$

which proves that $Z_H$ also has the property $(*)$.

Finally, let us show that $\mathbb{E}(\|Z\|^{-\eta}) < C_{R,\eta}$. For positive random variables we have

$$\begin{aligned}
\mathbb{E}\big(\|Z\|^{-\eta}\big) &= \int_0^R r^{-\eta} f_{\|Z\|}(r)\, dr \\
&= R^\eta - \lim_{r\to 0} r^{-\eta} P\big(\|Z\| \leq r\big) + \eta \int_0^R r^{-\eta-1} P\big(\|Z\| \leq r\big)\, dr \\
&\leq R^\eta + C\eta \int_0^R r^{-\eta-1+\eta K} dr \leq C_{R,\eta}\ ,
\end{aligned}$$

since $K > 1$ and $\eta > 0$. This proves Lemma A.1 and thus (26). □

To prove that $\gamma_d(y)$ is integrable on any bounded set, we apply the coarea formula to (64) to $g(x) = |J_K\Phi x|^{-1}\, 1_{\mathcal{A}}(\Phi x)$ where $\mathcal{A}$ is bounded:

$$\int_{\mathbb{R}^d} 1_{\mathcal{A}}(\Phi x)\, dx = \int_{\mathcal{A}} \int_{\Phi^{-1}(z)} |J_K\Phi x|^{-1}\, d\mathcal{H}^{d-K}\, dz = \int_{\mathcal{A}} \gamma_d(z)\, dz\,.$$

If $\mathcal{A}$ is a compact set then by assumption (B) it follows immediately that

$$\int_{\mathbb{R}^d} 1_{\mathcal{A}}(\Phi x)\, dx = \int_{\Phi^{-1}(\mathcal{A})} dx \le |B_{2,d}(C\sqrt{(d)})| < \infty, \tag{69}$$

which proves that $\gamma_d$ is integrable on a compact.

**A.2. Proof of Part (ii).** Let us now prove that for each $d$, $\gamma_d(y)$ can only vanish when $\mathrm{dist}(y, \overline{\Phi_d(\mathbb{R}^d)}) \le c/d$ for some fixed constant $c$. We will exploit the relationship between the sets $\Phi_d(\mathbb{R}^d)$ and $\Phi_{d/2}(\mathbb{R}^{d/2})$ thanks to the fact that $\Phi_d$ is an average potential over the domain.

The inequality (27) proves that $\gamma_d(y) = 0$ only if $\int_{\Phi^{-1}(y)} d\mathcal{H}^{d-K} = 0$. Since in finite integer dimensions the Hausdorff measure $\mathcal{H}^\ell$ is a multiple of the Lebesgue measure in $\mathbb{R}^\ell$, it is sufficient to show that whenever $y \in (\Phi_d(\mathbb{R}^d))^\circ$, the set $\Phi^{-1}(y)$ has positive Lebesgue measure of dimension $d-K$.

Without loss of generality, assume that $\Phi = (\phi_1, \ldots, \phi_K)$ are linearly independent functions. Otherwise, if there were a linear dependency of the form

$$\sum_{k \le K} \alpha_k \phi_k(x) \equiv 0\,,$$

then $\Phi_d(\mathbb{R}^d) = \partial\Phi_d(\mathbb{R}^d)$, thus $\Phi_d(\mathbb{R}^d)^o$ is empty and there is nothing to prove.

Let us write $d = r^\ell$, with $r$ denoting the length of the cube $\Lambda_d$. Suppose first that $r$ is even. Given $y \in (\Phi_{2^{-\ell}d}(\mathbb{R}^{2^{-\ell}d}))^\circ$ we will see that there exists $x \in \Phi^{-1}(y)$ whose Jacobian $J\Phi(x)$ has rank $K$. Then, by the Implicit Function Theorem, one can find a local reparametrization of $\Phi^{-1}(y)$ in a small neighborhood $V$ of the form $x = (v, \varphi(v))$ such that

$$\left\{(v, \varphi(v)); v \in V \subset \mathbb{R}^{d-K}, \varphi : V \to \mathbb{R}^K\right\} = \left\{(v, v') \in V \times \varphi(V);\ \Phi(v, v') = y\right\},$$

which has positive Lebesgue measure of dimension $d-K$.

Suppose first that $\Delta = 1$. Then the sets $\mathcal{S}_d = \Phi_d(\mathbb{R}^d) \subset \mathbb{R}^K$ satisfy $\mathcal{S}_d \subseteq \mathcal{S}_{q^\ell d}$ for $q = 1, 2, \ldots$. Indeed, given $y \in \mathcal{S}_d$, by definition there exists $x \in \mathbb{R}^d$ with $\Phi_d(x) = y$. Consider $\widetilde{x} = (x, \ldots, x)^{\otimes \ell} \in \mathbb{R}^{q^\ell d}$, a tiling of $x$, $q$ times along each dimension. By construction, $\widetilde{x}$ satisfies $\Phi_{q^\ell d}(\widetilde{x}) = y$ and therefore $y \in \mathcal{S}_{q^\ell d}$.

Now, consider $y \in \mathcal{S}_d^\circ \subseteq \mathcal{S}_{2^\ell d}^\circ$. If $\Phi_d$ was a smooth $C^s$ map, with $s > d-K$, then by Sard's theorem, the image of critical points $\{x \in \mathbb{R}^d; |J\Phi_d(x)| < K\}$ has zero Lebesgue measure in $\mathcal{S}_d$. Although one can extend Sard's theorem to weaker regularity assumptions [3], for our purposes we will use a weaker and simpler property that does not require the smoothness assumption, as described in the following lemma:

**Lemma A.2.** *Under the assumptions of the theorem, the set*

$$\mathcal{A} = \{y \in \mathbb{R}^K;\ 0 < \gamma_d(y) < \infty\}$$

*is dense in $\Phi_d(\mathbb{R}^d)$, and for each $y \in \mathcal{A}$ there exists $x \in \Phi_d^{-1}(y)$ with $|J\Phi_d(x)| > 0$.*

It follows that for a sufficiently small $\delta > 0$, a neighborhood $B(y, \delta) \subset \mathcal{S}_d$ of $y$ necessarily contains two points $y_1 = y + \eta$, $y_2 = y - \eta$ such that $\Phi_d^{-1}(y_1)$ or $\Phi_d^{-1}(y_2)$ contain a regular point. Let $x_1 \in \Phi_d^{-1}(y_1)$ and $x_2 \in \Phi_d^{-1}(y_2)$ be two points such that at least one is regular. The point $\widetilde{x} = (x_1^{\otimes \ell}, x_2^{\otimes \ell}) \in \mathbb{R}^{2^\ell d}$, obtained by concatenating $x_1$ and $x_2$ along the first coordinate, and tiling them along the rest, satisfies

$$\Phi_{2^\ell d}(\widetilde{x}) = \frac{1}{2}\big(\Phi_d(x_1) + \Phi_d(x_2)\big) = y\,,$$

and

$$|J\Phi_{2^\ell d}(\widetilde{x})| \geq \max\big(|J\Phi_d(x_1)|, |J\Phi_d(x_2)|\big) > 0\,,$$

which shows that we have just found an element $\widetilde{x}$ of $\Phi_{2^\ell d}^{-1}(y)$ with $\operatorname{rank}(J\Phi_{2^\ell d}(\widetilde{x})) = K$.

Suppose finally that $\Delta > 1$. The proof follows the same strategy, but we need to handle the border effect introduced by the support $\Delta$. In that case, given $y \in \mathcal{S}_d$, we consider $\widetilde{x} = (x, u, x)^{\otimes \ell}$, where $u$ has $2(\Delta - 1)$ zero coordinates and $x \in \Phi_d^{-1}(y)$. That is, we consider $2^\ell$ copies of $x$ separated by $2(\Delta - 1)$ zeroes along each dimension so that their potential functions do not interact.

Let $\widetilde{d} = (2r + 2(\Delta - 1))^\ell$. It follows that

$$\Phi_{\widetilde{d}}(\widetilde{x}) = \frac{2^\ell d \Phi_d(x)}{\widetilde{d}} = \left(1 + \frac{\Delta - 1}{r}\right)^{-\ell} \Phi_d(x) = \left(1 + \frac{\Delta - 1}{d^{1/\ell}}\right)^{-\ell} y\,,$$

which shows that $\operatorname{dist}(y; \mathcal{S}_{\widetilde{d}}) \lesssim C\ell\|y\|/d^{1/\ell}$ for any $y \in \mathcal{S}_d$.

Now consider $y$ in the open set $C_d = \mathcal{S}_d \cap \mathcal{S}_{\widetilde{d}}$, such that $\operatorname{dist}(y, \partial\mathcal{S}_d) \geq \|y\|\ell\Delta d^{-1/\ell}$. It follows from the previous argument that there exists small $\delta > 0$ and $x_1 \in \Phi_d^{-1}(y_1)$ with $|J\Phi_d(x_1)| > 0$ and $y_1 \in B(y, \delta) \cap \mathcal{S}_d \cap \mathcal{S}_{\widetilde{d}}$. We verify from the assumption that

$$y_2 = 2\left(1 + \frac{\Delta - 1}{r}\right)^{\ell} y - y_1 \in \mathcal{S}_d\,,$$

and therefore for any $x_2 \in \Phi_d^{-1}(y_2)$ the point $\widetilde{x} = (x_1; u, x_2)^{\otimes \ell}$ that contains $2^{\ell-1}$ copies of $x_1$ and $2^{\ell-1}$ copies of $x_2$ satisfies by construction

$$\Phi(\widetilde{x}) = \frac{d2^{\ell-1}y_1 + d2^{\ell-1}y_2}{\widetilde{d}} = y$$

and $\operatorname{rank}(J\Phi_{2d}(\tilde{x})) = K$. Finally, the case where $r$ is odd is treated analogously, but splitting the coordinates into $\lfloor \frac{r}{2} \rfloor$ and $\lceil \frac{r}{2} \rceil$ parts.

It remains to prove Lemma A.2. We know from part (i) that thanks to the coarea formula,

$$\forall \epsilon \ \forall y \in \left(\Phi_d(\mathbb{R}^d)\right)^\circ, \ 0 < \int_{\|z-y\|\le\epsilon} \gamma_d(z)\, dz = \int_{\|\Phi(x)-y\|\le\epsilon} dx < \infty .$$

It follows that $\mathcal{A} = \{z;\ 0 < \gamma_d(z) < \infty\}$ is dense in $\Phi_d(\mathbb{R}^d)$. But if $y \in \mathcal{A}$, by definition this implies that $\Phi_d^{-1}(y)$ has positive $(d-K)$-Hausdorff measure, and that there is necessarily $x \in \Phi_d^{-1}(y)$ with $|J\Phi_d(x)|^{-1} < \infty$, therefore with a full-rank Jacobian. □

**A.3. Proof of Part (iii).** In order to prove (28), we will again exploit the relationships between the sets $\mathcal{S}_d = \Phi_d(\mathbb{R}^d)$ as $d$ grows. We also first establish the result for $\Delta = 1$, and then generalize it to $\Delta > 1$. Denote $F_{d,\epsilon} = d^{-1} H(\mu^{\mathrm{mi}}_{d,\epsilon})$ the entropy rate associated with $y$ and $\epsilon$ and $\Omega_{d,\epsilon}(y) = \{x\,;\|\Phi_d(x) - y\| \le \epsilon\}$.

In the last section we proved that when $\Delta = 1$, $\mathcal{S}_d \subseteq S_{q^\ell d}$ for $q = 1, 2, \ldots$. For any $\epsilon > 0$ and $y \in \mathcal{S}_d$, observe that

$$\Omega_{d,\epsilon}(y) \underbrace{\otimes \cdots \otimes}_{2^\ell \text{ times}} \Omega_{d,\epsilon}(y) \subseteq \Omega_{2^\ell d,\epsilon}(y) . \tag{70}$$

Indeed, if $x \in \Omega_{d,\epsilon}(y) \underbrace{\otimes \cdots \otimes}_{2^\ell \text{ times}} \Omega_{d,\epsilon}(y)$, then by definition $x = (x_1, \ldots, x_{2^\ell})$ with

$$\|\Phi_d(x_i) - y\| \le \epsilon .$$

But

$$\Phi_{2^\ell d}(x) = 2^{-\ell} \sum_{i=1}^{2^\ell} \Phi_d(x_i)$$

and $\|\Phi_{2^\ell d}(x) - y\| \le \epsilon$ by the convexity of the $\mathbf{l}^2$ norm, thus $x \in \Omega_{2^\ell d,\epsilon}(y)$. It follows that

$$F_{2^\ell d,\epsilon} = d^{-1} 2^{-\ell} H\big(\mu^{\mathrm{mi}}_{2^\ell d,\epsilon}\big) \ge d^{-1} 2^{-\ell} \log \left( \left[ \int_{\|\Phi_d(x)-y\|\le\epsilon} dx \right]^{2^\ell} \right) = F_{d,\epsilon} . \tag{71}$$

Thus, for any fixed $d_0$, $y \in \mathcal{S}_{d_0}$ and $\epsilon > 0$, the sequence $F_k = F_{2^{k\ell} d_0,\epsilon}$ is increasing. Also, thanks to assumption (B), we have that

$$\forall\, d \ , \ x \in \Omega_{d,\epsilon}(y) \implies \|x\| \le C\sqrt{d}\big(\|y\| + \epsilon\big) ,$$

which implies that $|\Omega_{d,\epsilon}(y)| \leq |B_d(\sqrt{d}R_0)|$. Therefore

$$\forall\, d\ ,\ F^y_{d,\epsilon} \leq d^{-1} \log |B_d(\sqrt{d}R_0)|\,,$$

and we verify from $|B_d(R)| = \frac{\pi^{d/2}}{\Gamma(d/2+1)} R^d$ that $|B_d(\sqrt{d}R_0)| \simeq \widetilde{K}^d$ with $\widetilde{K} = 2\pi R_0^2 e$, which shows that $\lim_{d\to\infty} d^{-1}\log|B_d(\sqrt{d}R_0)| = \log \widetilde{K}$ and thus that the entropy rate $F_k$ is also upper bounded, and therefore its limit exists $\lim_{k\to\infty} F_k = \bar{F}$. We shall see later that the limit does not depend upon the choice of $d_0$.

Let us now prove the case when $\Delta > 1$. The idea is to show that (70) is now valid up to an error that becomes small as $d$ increases, provided that the potential $U$ is Holder continuous.

Consider $y \in \mathcal{S}_d$. Given $\epsilon > 0$, we form

$$\Psi_{2^\ell d,\epsilon}(y) = \big(\Omega_{d,\epsilon}(y)\big)^{\otimes 2^\ell}$$

as the Cartesian product of $2^\ell$ copies of $\Omega_{d,\epsilon}(y)$. When $\Delta = 1$, we just saw that

$$\Psi_{2^\ell d,\epsilon}(y) \subseteq \Omega_{2d,\widetilde{\epsilon}}(y) \tag{72}$$

with $\epsilon = \widetilde{\epsilon}$, but when $\Delta > 1$, let us see how to increase $\widetilde{\epsilon}$ so that (72) is verified. Given $x \in \Psi_{2^\ell d,\epsilon}(y)$, we write $x = (x_1, \ldots, x_{2^\ell})$ to denote its projections into each of the $2^\ell$ subdomains $C_{1,d}, \ldots, C_{2^\ell,d}$ of size $d$. We have

$$\begin{aligned}\Phi_{2^\ell d}(x) &= \frac{\sum_n Ux(n)}{2^\ell d}\\ &= 2^{-\ell}\sum_{k=1}^{2^\ell} d^{-1}\Bigg(\sum_{n\in C^\circ_{k,d}} Ux(n) + \sum_{n\in\partial C_{k,d}} Ux(n)\Bigg)\,,\end{aligned} \tag{73}$$

where each $C^\circ_{k,d}$ contains the interior of the domain that does not interact with the other domains, and $\partial C_{k,d} = C_{k,d} \setminus C^\circ_{k,d}$. We have $|\partial C_{k,d}| = d - (d^{1/\ell} - 2\Delta)^\ell$, thus

$$d^{-1}|\partial C_{k,d}| = 1 - \big(1 - 2\Delta d^{-1/\ell}\big)^\ell \lesssim \frac{\ell\Delta}{d^{1/\ell}}\,. \tag{74}$$

Since $|Ux(n)| \leq B\|x\|^\alpha$ with $\alpha < 2/\ell$ by the Holder assumption, and $\|x\| \leq C\sqrt{d}$ by assumption (B), we have $|Ux(n)| \leq B'd^{\alpha/2}$. It follows from (73) and (74) that

$$\begin{aligned}
\|\Phi_{2^\ell d}(x) - y\| &= \left\| 2^{-\ell} \sum_{k=1}^{2^\ell} \left[ d^{-1} \Big( \sum_{n \in C^\circ_{k,d}} Ux(n) + \sum_{n \in \partial C_{k,d}} Ux(n) \Big) - y \right] \right\| \\
&\leq 2^{-\ell} \sum_{k=1}^{2^\ell} \Big( \|\Phi_d(x_i) - y\| + 2B'd^{\alpha/2}\big(1 - \big(1 - 2\Delta d^{-1/\ell}\big)^\ell\big) \Big) \\
&\leq \epsilon + o\big(d^{\frac{\alpha}{2} - \frac{1}{\ell}} \ell \Delta\big) ,
\end{aligned}$$

Thus by taking $\widetilde{\epsilon} = \epsilon + o(d^{\frac{\alpha}{2} - \frac{1}{\ell}} \ell\Delta)$ (72) is verified. By denoting $\nu = \frac{\alpha}{2} - \frac{1}{\ell}$, it follows that the entropy rate $F_{d,\epsilon}$ satisfies

$$F_{d,\epsilon} \leq F_{2^\ell d, \epsilon + \widetilde{\ell} d^\nu} ,$$

with $\widetilde{\ell} = C\Delta\ell$, and $\nu < 0$ since $\alpha < 2/\ell$. By repeating the inequality for sufficiently large $d$ and $k = 1, 2, \ldots$ and $\epsilon > 0$ we have

$$F_{d,\epsilon} \leq F_{d2^{k\ell},\ \epsilon + \widetilde{\ell} d^\nu \sum_{k'=0}^{k} 2^{k'\ell\nu}} \leq F_{d2^{k\ell}, 2\epsilon} \leq \widetilde{C} , \tag{75}$$

and thus by defining

$$F_{\infty,\epsilon} := \lim_{k \to \infty} F_{d_0 2^{k\ell}, \epsilon_k} , \quad \text{with } \epsilon_k = \epsilon + \widetilde{\ell} d_0^\nu \sum_{k'=0}^{k} 2^{\ell\nu k'} \tag{76}$$

we have shown that its entropy rate is well-defined for each $\epsilon > 0$ and $d_0$ sufficiently large.

It remains to be shown that this limit does not depend upon $d_0$. Suppose $F_{\infty,\epsilon,0} \neq F_{\infty,\epsilon,1}$ where $F_0$ is associated with $d_0$ and $F_1$ is associated with $d_1$, and suppose $d_1 > d_0$ without loss of generality. Let $r_i = d_i^{1/\ell}$ for $i = 0, 1$.

Observe that an analogous argument to (73) shows that if $r = r_a + r_b$, then

$$F_{r^\ell, \widetilde{\epsilon}} \geq \frac{r_a}{r} F_{r_a^\ell, \epsilon} + \frac{r_b}{r} F_{r_b^\ell, \epsilon} , \tag{77}$$

and

$$F_{l^\ell d, \widetilde{\epsilon}} \geq F_{d,\epsilon} \quad \text{for } l = 1, 2, \ldots , \tag{78}$$

with $\widetilde{\epsilon} = \epsilon + o\,(d^\nu \ell\Delta)$. Consider now large integers $k$ and $\widetilde{k} \simeq \sqrt{k}$, and let $q, \widetilde{q}$ denote respectively the quotient and residual such that

$$r_1 2^k = r_0 2^{\widetilde{k}} q + \widetilde{q}$$

with $0 \leq \widetilde{q} < r_0 2^{\widetilde{k}}$. Then, for any $\delta > 0$, by choosing $k$ large enough we obtain from (77) and (78) that

$$|F_{d_1 2^k \ell, \widetilde{\epsilon}} - F_1| \leq \delta/4\,,$$
$$|F_{d_0 2^{\ell \widetilde{k}} q^\ell, \widetilde{\epsilon}} - F_0| \leq \delta/4\,,$$

and
$$|F_{d_1 2^k \ell, \overline{\epsilon}} - F_{d_0 2^{\widetilde{k}\ell} q^\ell, \widetilde{\epsilon}}| \leq \delta/4\,, \tag{79}$$

with $\overline{\epsilon} = \widetilde{\epsilon} + o(d^\nu \ell \Delta)$.

Finally, let us show that $F_{d,\epsilon}$ is continuous with respect to $\epsilon$ for $\epsilon > 0$. Let us denote $\gamma_{d,\epsilon} = \int_{\|z-y\|\leq\epsilon} \gamma_d(z)\,dz$. Since $F_{d,\epsilon} = d^{-1}\log(\gamma^y_{d,\epsilon})$ and $\gamma_d(y) > 0$ for all $y \in \mathcal{S}^\circ_d$ from the previous section, it is sufficient to show that $\gamma_{d,\epsilon}$ is continuous with respect to $\epsilon$. Let $\widetilde{\epsilon} = \epsilon + \delta$ with $\epsilon > 0$, and suppose $\delta > 0$ without loss of generality. By denoting $Q(\delta, \epsilon, y) = \{z\,;\, \epsilon < \|z - y\| \leq \epsilon + \delta\}$, we have

$$\begin{aligned}|\gamma_{d,\widetilde{\epsilon}} - \gamma_{d,\epsilon}| &= \int_{\epsilon<\|z-y\|\leq\epsilon+\delta} \gamma_d(z)\,dz = \int \gamma_d(z)\mathbf{1}_{Q(\delta,\epsilon,y)}(z)\,dz \\ &:= \int \gamma_{d\,\delta}(z)\,dz\end{aligned}$$

For each $z$, $\gamma_{d\,\delta}(z) = \gamma_d(z)\mathbf{1}_{Q(\delta,\epsilon,y)}(z)$ converges pointwise to 0 as $\delta \to 0$, except for a set of measure zero, $\{z; \|z - y\| = \epsilon\}$. Also, $|\gamma_{d\,\delta}| \leq \gamma_d$, which is integrable in $\Phi(\Omega_d)$ by part (i). We can thus apply the dominated convergence theorem, and conclude that

$$\lim_{\delta\to 0} \int \gamma_{d\,\delta}(z)\,dz = \int \big(\lim_{\delta\to 0} \gamma_{d\,\delta}(z)\big)\,dz = 0\,,$$

which shows that $\gamma_{d,\epsilon}$ is continuous with respect to $\epsilon$.

It follows from (79) that

$$|F_{d_1 2^k \ell, \widetilde{\epsilon}} - F_{d_0 2^{\widetilde{k}\ell} q^\ell, \widetilde{\epsilon}}| \to 0 \text{ as } k \to \infty\,,$$

but $F_{d_1 2^k \ell, \widetilde{\epsilon}} \to F_1$ and $F_{d_0 2^{\widetilde{k}\ell} q^\ell, \widetilde{\epsilon}} \to F_0$ as $k \to \infty$, which is a contradiction with the fact that $F_0 \neq F_1$. □

## B. Proof of Corollary 3.2

We saw in Theorem 3.7 that the entropy rate of the microcanonical measure can be measured with the co-area formula as $d^{-1}H(\mu^{\text{mi}}_{d,\epsilon}) = d^{-1}\log\int_{\|z-y\|\leq\epsilon}\gamma_d(z)\,dz$ and that $\gamma_d(z) > 0$ in the interior of $\Phi_d(\mathbb{R}^d)$. As $\epsilon \to 0$, we can interpret the previous formula in terms of an $L^1(\mathbb{R}^K)$ approximate identity $h_\epsilon(z) = C_K\epsilon^{-K}\mathbf{1}_{\|z\|\leq\epsilon}(z)$:

$$C_K\epsilon^{-K}\int_{\|z-y\|\leq\epsilon}\gamma_d(z)\,dz = \gamma_d \star h_\epsilon(y) \to \gamma_d(y) \text{ as } \epsilon \to 0$$

in $L^1(\mathbb{R}^K)$. One can verify that, by possibly reparametrising $\epsilon$, this implies pointwise convergence for almost every $y$, so

$$\left|\log\left(C_K \epsilon^{-K} \gamma^y_{d,\epsilon}\right) - \log \gamma_d(y)\right| \underset{\epsilon\to 0}{\to} 0\,,\ a.e.\,, \tag{80}$$

which shows that $d^{-1}H(p^y_{d,\epsilon}) \simeq \frac{-K}{d}\log\epsilon$ as $\epsilon \to 0$. □

## C. Proof of Proposition 3.3

Properties (A) and (B) are verified for (i)–(ii) because the potentials $U$ are continuous and the resulting features $\Phi$ always include $d^{-1}\|x\|^2$ respectively. We thus focus on proving property (C).

Part (i) is easily obtained, since the $\mathbf{l}^2$ wavelet model has a Jacobian $J\Phi(x)$ that is linear with respect to $x$, and therefore it has absolutely continuous density relative to the Lebesgue measure.

Part (ii) is proved by directly controlling $|J\Phi_d(x)|^{-1}$. A direct computation shows that $|J\Phi_d(x)| = d^{-1}\sqrt{d\|x\|^2 - \|x\|_1^2}$, which only vanishes when $|x|$ is a constant vector. Therefore, for $y \neq (\alpha, \Lambda_d \alpha)$, $\Phi^{-1}_{d,\epsilon}(y)$ does not contain those points for sufficiently small $\epsilon$.

Let us now show part (iii). The Jacobian matrix in that case is given by

$$J\Phi_d(x)_j = d^{-1}\mathrm{Re}\left\{\left(\frac{x \star h_j}{|x \star h_j|}\right) \star h_j^*\right\},$$

with $j \leq K$. We proceed by induction over the scale $K$. Suppose first $K = 1$. Since $h_j$ has compact spatial support, its Fourier transform only contains a discrete number of zeros. Denote by $\Delta_j$ the spatial support of $h_j$. We can thus generate all but a zero-measure set of unitary signals $z$ with $z_s = e^{i\theta_s}$, $s = 1, \ldots, \Delta_j$ from the uniform measure over $x$ using $z = \frac{x \star h_j}{|x \star h_j|}$. In the uniform phase space defined by $\theta_1, \ldots, \theta_{\Delta_j}$, the event $|\det \bar{J} U(\bar{X}_1)| \leq y$ has a probability proportional to $y$, since it is equivalent to

$$\left|\sum_s \cos(\theta_s)\mathrm{Re}(h_j^*(s)) - \sum_s \sin(\theta_s)\mathrm{Im}(h_j^*(s))\right| \leq y\,.$$

Suppose now the result holds for the $K-1$ filters in the family with smallest spatial support, and let us show how to extend it to an extra filter $h_K$ with strictly larger spatial support. Among the variables $\bar{X} \in \mathbb{R}^{2\Delta+1}$, a subset of them, say $R_K$, only affect the $K$-th output corresponding to filter $h_K$. It follows that a set $S \subset \mathbb{R}^K$ with shrinking measure necessarily introduces constraints on the variables in $R_K$, and therefore $P(Z \in S) \leq |S|^{1/K}$. □

## D. Proof of Theorem 3.4

(i) Let us first prove that volume preserving symmetries of $\Phi_d(x)$ are symmetries of the microcanonical maximum entropy measure. If for all $x \in \mathbb{R}^d$, $\Phi_d(L^{-1}x) = \Phi_d(x)$ then a microcanonical set $\Omega_{d,\epsilon}$ is invariant to the action of $L$ and $L^{-1}$. Since $L$ preserves volume and hence the Lebesgue measure of a set, for any measurable set $\mathcal{A}$, since $\mu^{\mathrm{mi}}_{d,\epsilon}$ is supported over $\Omega_{d,\epsilon}$ and uniform relatively to the Lebesgue measure, we have

$$\begin{aligned}\mu^{\mathrm{mi}}_{d,\epsilon}\big[L^{-1}\mathcal{A}\big] &= \mu^{\mathrm{mi}}_{d,\epsilon}\big[L^{-1}\mathcal{A} \cap \Omega_{d,\epsilon}\big] \\ &= \mu^{\mathrm{mi}}_{d,\epsilon}\big[L^{-1}\big(\mathcal{A} \cap \Omega_{d,\epsilon}\big)\big] \\ &= \mu^{\mathrm{mi}}_{d,\epsilon}\big[\mathcal{A} \cap \Omega_{d,\epsilon}\big] = \mu^{\mathrm{mi}}_{d,\epsilon}[\mathcal{A}],\end{aligned}$$

so $L$ is a symmetry of $\mu^{\mathrm{mi}}_{d,\epsilon}$.

(ii) We prove that symmetries of $\Phi_d(x)$ and $\mu_0$ are symmetries of $\mu_n$, by induction on $n$. It is trivially valid for $n = 0$. Suppose now by induction that $\mu_n$ is invariant to the action of $L$ which is a symmetry of $\Phi_d$. From (31), $\mu_{n+1} = \varphi_{n,\#}\mu_n$, with

$$\varphi_n(x) = x - \kappa_n J\Phi_d(x)^\top \big(\Phi_d(x) - y\big)\,.$$

Let us verify that $\varphi_n$ is equivariant to the action of $L$: $\varphi_n L^{-1}x = L^{-1}\varphi_n x$ for all $x$. Since $\Phi_d(L^{-1}x) = \Phi_d(x)$, and since $L$ is linear

$$J\Phi_d(L^{-1}x)^\top = L^{-1}\big(J\Phi_d(L^{-1}x)\big)^\top = L^{-1}\big(J\Phi_d(x)\big)^\top \tag{81}$$

so

$$\begin{aligned}\varphi_n L^{-1}x &= L^{-1}x - \kappa_n J\Phi_d(L^{-1}x)^\top\big(\Phi_d(L^{-1}x) - y\big) \\ &= L^{-1}x - L^{-1}\kappa_n J\Phi_d(x)^\top\big(\Phi_d(x) - y\big) \\ &= L^{-1}\varphi_n x\,,\end{aligned}$$

which proves that $\varphi_n$ is equivariant to the action of $L$. Moreover, if $\varphi_n$ is equivariant to the action of $L$ then we verify that it is equivariant to the action of $L^{-1}$. Also, observe that

$$\begin{aligned}\varphi_n^{-1}(L^{-1}(\mathcal{A})) &= \big\{x; \varphi_n(x) \in L^{-1}\mathcal{A}\big\} \\ &= \big\{x; L\varphi_n(x) \in \mathcal{A}\big\} \\ &= \big\{x; \varphi_n(Lx) \in \mathcal{A}\big\} \\ &= L^{-1}\varphi_n^{-1}(\mathcal{A})\,.\end{aligned}$$

Finally, using the definition of pushforward measure, $\mu_{n+1} = \varphi_{n,\#}\mu_n$, for any measurable $\mathcal{A}$, the induction hypothesis yields

$$\begin{aligned}\mu_{n+1}\big[L^{-1}\mathcal{A}\big] &= \mu_n\big[\varphi_n^{-1}(L^{-1}\mathcal{A})\big]\\ &= \mu_n\big[L^{-1}\varphi_n^{-1}(\mathcal{A})\big]\\ &= \mu_n\big[\varphi_n^{-1}(\mathcal{A})\big] = \mu_{n+1}[\mathcal{A}]\,,\end{aligned}$$

which proves that $\mu_{n+1}$ is also invariant to the action of $L$.

(iii) We prove that an orthogonal operator which preserves a stationary mean is a symmetry of a Gaussian measure $\mu_0$ of $d$ i.i.d Gaussian random variables. Applying the statement (ii) then implies the statement (iii). Let $m_0$ be the mean of each of the $d$ Gaussian random variables. The Gaussian measure $\mu_0$ is uniform over all spheres of $\mathbb{R}^d$ centered over the stationary mean $m_0\,\mathbf{1}$. An orthogonal operator $L$ which preserves the stationary mean leaves invariant all spheres centered in $m_0\,\mathbf{1} \in \mathbb{R}^d$. Indeed $L(m_0\,\mathbf{1}) = m_0\,\mathbf{1}$ and $\|Lx\|^2 = \|x\|^2$ so

$$\|Lx - m_0\mathbf{1}\|^2 = \|L(x - m_0\mathbf{1})\|^2 = \|x - m_0\mathbf{1}\|^2.$$

If $S(m\mathbf{1}, r)$ is a sphere centered in $m\mathbf{1}$ of radius $r$ then $\mathbb{R}^d = \cup_{(m,r)\in\mathbb{R}\times\mathbb{R}^+} S(m\mathbf{1}, r)$. So for any measurable set $\mathcal{A}$

$$\begin{aligned}\mu_0[L^{-1}\mathcal{A}] &= \mu_0\big[L^{-1}\mathcal{A} \cap \cup_{(m,r)\in\mathbb{R}\times\mathbb{R}^+} S(m\mathbf{1}, r)\big]\\ &= \mu_0\big[\cup_{(m,r)\in\mathbb{R}\times\mathbb{R}^+} L^{-1}\big(\mathcal{A} \cap S(m\mathbf{1}, r)\big)\big]\\ &= \mu_0\big[\cup_{(m,r)\in\mathbb{R}\times\mathbb{R}^+} \mathcal{A} \cap S(m\mathbf{1}, r)\big] = \mu_0[\mathcal{A}],\end{aligned}$$

so $L$ is a symmetry of $\mu_0$. □

## E. Proof of Theorem 3.7

**E.1. Proof of Part (i).** Let us first show how the strict saddle condition (33) implies that the minimisation $\mathcal{E}(x)$ has no poor local minima. The statement follows directly from [31], which shows that when the saddle points are strict, gradient descent does not converge to those saddle points, up to a set of initialization values with Lebesgue measure 0. Observe first that $\kappa_n < \eta^{-1}$ ensures that $\varphi_n(x) = x - \eta\nabla E(x)$ is a diffeomorphism for each $n$. Observe also that a critical point $x$ such that $\nabla E(x) = J\Phi_d(x)^T(\Phi_d(x) - y) = 0$ necessarily falls into two categories. Either $\Phi_d(x) = y$, which implies that $x$ is a global optimum, or $x$ is such that $J\Phi_d(x)^T v = 0$ with $v = \Phi_d(x) - y \neq 0$. We verify that assumption (33) implies that in that case $x$ is a strict saddle point by observing that the Hessian of $E$ satisfies

$$\nabla^2 E(x) = \sum_{k=1}^{K} \nabla^2\Phi_k(x)v_k + J\Phi(x)^T J\Phi(x)\,.$$

Since $\mu_0$ is absolutely continuous with respect to the Lebesgue measure, we can apply Theorem 2.1 from [38], and establish that gradient descent does not converge to any saddle point with probability 1.

Let us now prove that the hypothesis that $|J\Phi_d(x)| > 0$ for $x \in \Phi_d^{-1}(y)$ with $y \in \Phi_d(\mathbb{R}^d)^\circ$, together with the strict saddle condition, implies that the gradient descent sequence $x_n$ has a limit $\lim_{n\to\infty} x_n$ (that may depend upon $x_0$). For that, we will apply the following result from [1]:

**Theorem E.1.** *If $E(x)$ is twice differentiable, has compact sub-level sets, and the Hessian $\nabla^2 E(x)$ is non-degenerate on the normal space to the level set of local minimisers, then $x_n$ has a limit, denoted $x_\infty := \lim_{n\to\infty} x_n$.*

Indeed, since $\Phi_d$ satisfies assumption (B), it follows that the sub-level sets of $E$, $\{x; E(x) \le t\}$ are compact for each $t$. We need to show that the Hessian of $E$ is non-degenerate on the normal space of $\Phi_d^{-1}(y)$. Since $\gamma_d > 0$ for $y \in \Phi_d(\mathbb{R}^d)^\circ$ for sufficiently large $d$ from Theorem 3.1, $\Phi_d^{-1}(y)$ has positive $d-K$-dimensional Hausdorff measure, hence it is sufficient to show that $\nabla^2 E(x)$ has $K$ strictly positive eigenvalues when $x \in \Phi^{-1}(y)$. But by definition,

$$\nabla^2 E(x) = \sum_{k\le K} \nabla^2 \phi_k(x)\big(\phi_k(x) - y_k\big) + J\Phi_d(x)^T J\Phi_d(x)\,,$$

thus

$$\nabla^2 E(x) = J\Phi_d(x)^T J\Phi_d(x) \text{ for } x \in \Phi_d^{-1}(y)\,. \tag{82}$$

Therefore, if $|J\Phi_d(x)| > 0$ for $x \in \Phi_d^{-1}(y)$, we can apply Theorem E.1, and conclude that the iterates $x_n$ from gradient descent have a limit, for each $x_0 \sim \mu_0$.

We have just proved that

$$\mathrm{P}_{\mu_0}\big\{(x_n)_n \text{ is Cauchy}\big\} = 1\,,$$

or, equivalently, that $X_n \sim \mu_n$ is almost surely Cauchy, which implies [42] that $\mu_n$ converges almost surely to a certain measure $\mu_\infty$. Moreover, since $\lim_{n\to\infty} \|\nabla E(x_n)\| = 0$, the strict saddle condition implies that $x_n$ does not converge to saddle points, so we conclude that necessarily

$$\mu_\infty\big[\Phi_d^{-1}(y)\big] = \mathrm{P}_{\mu_0}\big\{\lim_{n\to\infty} x_n \in \Phi_d^{-1}(y)\big\} = 1\,,$$

therefore that $\mu_\infty$ is supported in the microcanonical ensemble $\Phi_d^{-1}(y)$, which finishes the proof. $\square$

**E.2. Proof of Part (ii).** We first compute how the entropy is modified at each gradient step. By definition of the pushforward measure, for any diffeomorphism $\varphi$ and any measurable $g$

$$\mathbb{E}_{x\sim\varphi_\# \mu}\, g(x) = \mathbb{E}_{x\sim\mu}\, g(\varphi(x))\,.$$

Also, from a change of variables we have, by denoting $\widetilde{\mu} = \varphi_{\#}\mu$, $\widetilde{\mu}(x) = |J\varphi^{-1}(x)|\mu(\varphi^{-1}(x))$ , and thus

$$\log \widetilde{\mu}(x) = \log \mu\big(\varphi^{-1}(x)\big) - \log \big|J\varphi\big(\varphi^{-1}(x)\big)\big| .$$

It follows that

$$-\mathbb{E}_{x\sim\widetilde{\mu}} \log \widetilde{\mu}(x) = -\mathbb{E}_{x\sim\mu} \log \mu(x) + \mathbb{E}_{x\sim\mu} \log |J\varphi(x)|$$

and hence

$$H(\varphi_{\#}\mu) = H(\mu) - \mathbb{E}_{\mu} \log |J\varphi(x)| . \tag{83}$$

The change in entropy by applying the diffeomorphism is thus given by the term $\mathbb{E}_{\mu} \log |J\varphi(x)|$, and thus the entropy of $\mu_n$ is given by

$$H(\mu_n) = H(\mu_0) - \sum_{n' \le n} \mathbb{E}_{\mu_{n'}} \log |J\varphi_n(x)|$$

By definition, the Jacobian of $\varphi_n$ is

$$J\varphi_n(x) = \mathbf{1} - \gamma_n \Bigg( \sum_{k \le K} \nabla^2 \phi_k(x)\big(\phi_k(x) - y_k\big) + J\Phi_d(x)^T J\Phi_d(x) \Bigg) . \tag{84}$$

We know that $\Phi$ is Lipschitz, which implies that $\|J\Phi(x)\| \le \beta$, and that $\nabla\Phi$ is also Lipschitz, meaning that $\|\nabla^2\phi_k(x)\| \le \eta$ for all $k$. Applying the Cauchy–Schwartz inequality, it follows that

$$\Big\| \sum_{k \le K} \nabla^2 \phi_k(x)\big(\phi_k(x) - y_k\big) \Big\| \le \eta K \sqrt{E(x)} .$$

We abuse notation and redefine $\eta := \eta K$ since $K$ is a constant. Also, the term $J\Phi(x)^T J\Phi(x)$ is of rank at most $K$. We can thus write $J\varphi_n(x)$ as

$$J\varphi_n(x) = A_n(x) + B_n(x) , \tag{85}$$

with $A_n(x)$ full rank $d$ and with singular values within the interval

$$\big(1 - \gamma_n \eta \sqrt{E(x)}, 1 + \gamma_n \eta \sqrt{E(x)}\big);$$

and $-B_n(x)$ positive semidefinite of rank $K$, with singular values bounded by $\gamma_n \beta^2$.

It follows that the singular values of $J\varphi_n(x)$, called $\lambda_1, \dots, \lambda_d$, satisfy

$$
\begin{aligned}
|\log|J\varphi_n(x)|| &\leq \sum_{i=1}^{d} |\log \lambda_i| \\
&\leq \sum_{i=1}^{d-K} \max\left(|\log(1+\gamma_n \eta \sqrt{E(x)})|, |\log\left(1-\gamma_n\eta\sqrt{E(x)}\right)|\right) \\
&\qquad\qquad\qquad + \sum_{i=1}^{K} |\log(1-\gamma_n \beta^2)| \\
&\leq (d-K)\log\left(1+\gamma_n\eta\sqrt{E(x)}\right) + K\log(1+\gamma_n\beta^2) + o(\gamma_n^2)
\end{aligned}
$$

and thus up to second order terms we have

$$
\begin{aligned}
\mathbb{E}_{\mu_n} \log|J\varphi_n(x)| &\leq (d-K)\log\left(1+\gamma_n\eta\mathbb{E}_{\mu_n}\sqrt{E(x)}\right) + K\log(1+\gamma_n\beta^2)\,, \\
&\leq (d-K)\gamma_n\eta\mathbb{E}_{\mu_n}\sqrt{E(x)} + K\gamma_n\beta^2\,,
\end{aligned} \tag{86}
$$

where we have used Jensen's inequality on the concave function $\log(1+x)$ and $\log(1+x) \leq x$ for $x \geq 0$ to obtain the inequality $\mathbb{E}\log(1+X) \leq \log(1+\mathbb{E}X)$. Denoting by $r_n = \mathbb{E}_{\mu_n}\sqrt{E(x)}$ the average distance to the microcanonical ensemble at iteration $n$, it results from (86) that after $n$ steps of gradient descent the entropy rate has decreased at most

$$
\left(1-\frac{K}{d}\right)\eta \sum_{n'\leq n} \gamma_{n'} r_{n'} + \frac{K}{d}\beta^2 \sum_{n'\leq n} \gamma_{n'}\,. \qquad \square
$$

## F. Proof of Corollary 3.8

The proof is a direct application of Theorem 3.7 and Sard's theorem, that states that if $\Phi_d$ is a $\mathbf{C}^\infty$ Lipschitz function, then the image of its critical points $\{x\,;\,|J\Phi_d(x)|=0\}$ has zero measure. We can thus apply Theorem E.1 from Part (ii) of the proof of Theorem 3.7 for almost every $y$. $\square$

## G. Proof of Theorem 3.9

We show that $\Phi_d(x) = \{d^{-1}\|x \star h_k\|_2^2\}_k$ satisfies the strict saddle condition. Here $x \in \mathbb{R}^d$, and we recall that the Fourier transform is defined as $\widehat{x}(\omega) = \sum_u x(u)e^{-i\omega u 2\pi/d}$, with $\omega \in (-d/2, d/2]$. The gradient of the loss function $E(x) = \frac{1}{2}\|\Phi(x) - y\|^2$ is

$$
\nabla E(x) = J\Phi_d(x)^T\left(\Phi_d(x) - y\right),
$$

and its Hessian is

$$\nabla^2 E(x) = \sum_k \nabla^2 \phi_k(x) v_k + J\Phi_d(x)^\top J\Phi_d(x)\,,$$

where $v_k = \phi_k(x) - y_k$. Expressing the gradient and the Hessian in the Fourier domain yields

$$\nabla E(\hat{x}) = \hat{x} \cdot \Big(\sum_k v_k |\hat{h}_k|^2\Big) \tag{87}$$

$$\nabla^2 E(\hat{x})(\omega, \omega') = \sum_k v_k |\hat{h}_k(\omega)|^2 \delta(\omega - \omega') + \hat{x}(\omega)|\hat{h}_k(\omega)|^2 \hat{x}(\omega')^* |\hat{h}_k(\omega')|^2\,. \tag{88}$$

The Hessian thus contains a diagonal term and a rank-$K$ term. We need to show that a critical point $x$ satisfying $\nabla E(\hat{x}) = 0$ with $\|v\| > 0$ has a Hessian matrix with at least one negative eigenvalue. From (87), it follows that a critical point satisfies

$$\forall \omega\ ,\ \hat{x}(\omega) \cdot \Big(\sum_k v_k |\hat{h}_k(\omega)|^2\Big) = 0\,. \tag{89}$$

Let $C = \{\omega\ ;\ \hat{x}(\omega) \neq 0\}$. The Hessian is expressed in terms of block matrices regrouping the frequencies in $C$ as

$$\nabla^2 E(\hat{x}) = \left(\begin{array}{c|c} \mathrm{M} & 0 \\ \hline 0 & \nabla^2_{C,C} \end{array}\right),$$

where M is the diagonal matrix of size $(d - |C|) \times (d - |C|)$ given by the frequencies outside $C$, such that $\hat{x}(\omega) = 0$:

$$\mathrm{M}_{\omega,\omega} = \sum_k v_k |\hat{h}_k(\omega)|^2\ ,\ \omega \notin C\,.$$

We examine the diagonal block corresponding to M. The image of $\Phi_d$ is the convex cone $\mathcal{C}$ in $\mathbb{R}_+^K$ determined by the directions $o_\omega = (|\hat{h}_1(\omega)|^2, \ldots, |\hat{h}_K(\omega_j)|^2) \in \mathbb{R}^K$, $\omega = 1, \ldots, d$. Without loss of generality, we assume here that $\|o_\omega\| > 0$ for all $\omega$, since frequencies that are invisible to all the filters do not play any role in the gradient descent. The target $y$ is by hypothesis in the interior of $\mathcal{C}$. Further, any two directions $o, o'$ in $\mathcal{C}$ satisfy

$$\langle o, o' \rangle = \sum_k |\hat{h}_k(\omega)|^2 |\hat{h}_k(\omega')|^2 > 0\,,$$

since the filters have compact spatial support.

If $C$ is empty, then $x = 0$, which implies that $v = \Phi(x) - y = -y$ has all its entries negative, and therefore $\mathrm{diag}(\sum_k v_k |\hat{h}_k(\omega)|^2) < 0$. We shall thus assume in the following that $C$ is non-empty. Similarly, we verify that the space spanned by $o_\omega$, $\omega \in C$, cannot have full rank $K$. Indeed, if this was the case, the first order optimality condition (89) reveals that $v$ should be orthogonal to all directions $o_\omega$, $\omega \in C$. Since this system has rank $K$, this contradicts the fact that $v \neq 0$.

We can thus write $\mathcal{C}$ as generated by directions $\mathcal{O}_C = \{o_\omega ; \omega \in C\}$ and $\mathcal{O}_{\bar{C}} = \{o_\omega ; \omega \notin C\}$, with $|\mathcal{O}_{\bar{C}}| > 0$, $|\mathcal{O}_C| > 0$. Since $y$ is in the interior, it follows that

$$y = \sum_{\omega \in C} \beta_\omega o_\omega + \sum_{\omega \notin C} \gamma_\omega o_\omega \ , \ \beta_\omega, \gamma_\omega > 0 \ \forall \ \omega \,. \tag{90}$$

We need to show that there exists at least one $\omega \notin C$ such that $\langle v, o_\omega \rangle < 0$. Suppose otherwise, i.e. that for all $\omega \notin C$, $\langle \Phi_d(x), o_\omega \rangle \geq \langle y, o_\omega \rangle$. Since $o_\omega \in \mathcal{O}_C \Rightarrow \langle \Phi_d(x), o_\omega \rangle = \langle y, o_\omega \rangle$ by the first order critical conditions, we have

$$\begin{aligned} \langle y, y \rangle &= \sum_{\omega \in C} \beta_\omega \langle o_\omega, y \rangle + \sum_{\omega \notin C} \gamma_\omega \langle o_\omega, y \rangle \\ &\leq \sum_{\omega \in C} \beta_\omega \langle o_\omega, \Phi_d(x) \rangle + \sum_{\omega \notin C} \gamma_\omega \langle o_\omega, \Phi_d(x) \rangle \,. \end{aligned} \tag{91}$$

On the other hand, from (90) we also have

$$\langle y, \Phi_d(x) \rangle = \sum_{\omega \in C} \beta_\omega \langle o_\omega, \Phi_d(x) \rangle + \sum_{\omega \notin C} \gamma_\omega \langle o_\omega, \Phi_d(x) \rangle \,, \tag{92}$$

and since $\Phi(x) = \sum_{\omega \in C} \alpha_\omega o_\omega$ is a linear combination of vectors in $\mathcal{O}_{\bar{C}}$, we also have $\langle \Phi(x), y \rangle = \langle \Phi(x), \Phi(x) \rangle$. This implies from (91) that

$$\langle y, y \rangle \leq \langle y, \Phi_d(x) \rangle = \langle \Phi_d(x), \Phi_d(x) \rangle \,, \tag{93}$$

which leads to $y = \Phi(x)$ and therefore $v = 0$, which is a contradiction.

Finally, if $x \in \Phi_d^{-1}(y)$ for $y \in \Phi_d(\mathbb{R}^d)^\circ$, then $y$ falls necessarily inside the convex hull of $\mathcal{C}$, which implies that $\{\nabla \phi_k(x) = \hat{x}(\omega) \cdot |\hat{h}_k|^2(\omega)\}_{k \leq K}$ have rank $K$. This concludes the proof. □

## H. Proof of Proposition 4.1

If $\gamma = 0$ then (39) proves that

$$\|x\|_2^2 = \|x \star \psi_{J,0}\|_2^2 + \sum_{j'=1}^{\log_2 d} \sum_q \|x \star \psi_{j',q'}\|_2^2.$$

If $J = \log_2 d$ then $\psi_{J,0}(u) = d^{-1} 1_{\Lambda_d}$ and $x \star \psi_{J,0}(u)$ is the average of $x$ over $\Lambda_d$. We thus get

$$\|x\|_2^2 = d^{-1}\Big(\sum_u x(u)\Big)^2 + \sum_{j'=1}^{\log_2 d} \sum_q \|x \star \psi_{j',q'}\|_2^2. \tag{94}$$

Replacing $x$ by $|x \star \psi_{j,q}|$ gives

$$\|x \star \psi_{j,q}\|_2^2 = d^{-1}\|x \star \psi_{j,q}\|_1^2 + \sum_{j'=1}^{\log_2 d} \sum_q \||x \star \psi_{j,q}| \star \psi_{j',q'}\|_2^2.$$

We finally prove (47) by decomposing each term $\||x \star \psi_{j,q}| \star \psi_{j',q'}\|_2^2$ into an $\mathbf{l}^1$ norm plus a sum of $\mathbf{l}^2$ norms, obtained replacing $x$ by $||x \star \psi_{j,q}| \star \psi_{j',q'}|$ in (94). □

## I. Proof of Theorem 5.1

Let us first prove property (i). Young's inequality is proved by observing that

$$\begin{aligned}\|x \star \psi_{j,q}\|_1 &= \sum_{n \in \Lambda_d} \Big| \sum_{u \in \Lambda_d} x(u)\, \psi_{j,q}(n-u)\Big| \\ &\leq \sum_{n \in \Lambda_d} \sum_{u \in \Lambda_d} |x(u)\, \psi_{j,q}(n-u)| = \|x\|_1\, \|\psi_{j,q}\|_1\,.\end{aligned}$$

The inequality is an equality if and only if for any fixed $n$, the product $x(u)\,\psi_{j,q}(n-u)$ has a constant phase when $u$ varies. Since $x(u)$ is real, its phase is either 0 or $\pi$. It implies that $\psi_{j,q}(n-u)$ has a phase modulo $\pi$ which does not depend upon $u$ when $x(u)\psi_{j,q}(n-u) \neq 0$ and hence $x(u) \neq 0$. Since the phase of $\psi$ is $\varphi(\xi.u)$, the phase of $\psi_{j,q}(u) = 2^{-\ell j}\psi(2^{-j} r_q^{-1} u)$ is $\varphi(2^{-j}\xi_q.u)$ with $\xi_q = r_q \xi$ so

$$\forall u \in \Lambda_d\ ,\ \varphi\big(2^{-j}\xi_q.(n-u)\big) = a(2^{-j}n) + k\pi \text{ if } x(u)\,\psi_{j,q}(n-u) \neq 0 \text{ with } k \in \mathbb{Z}\,. \tag{95}$$

Since $\varphi$ is bi-Lipschitz, there exists $\beta > 0$ such that

$$\beta^{-1}|a - a'| \leq |\varphi_q(a) - \varphi_q(a')| \leq \beta|a - a'|\,. \tag{96}$$

Since $\psi_q(0) \neq 0$ and $\psi_q$ is continuous, there exists $\alpha > 0$ such that $|\psi_q(u)| > 0$ for $u \in [-\alpha, \alpha]^\ell$. If $2^{-j}|u - u'| \leq 2\alpha$ then for $n = (u + u')/2$ we have $2^{-j}|n - u| \leq \alpha$ and $2^{-j}|n - u'| \leq \alpha$, so $\psi_{j,q}(n-u) \neq 0$ and $\psi_{j,q}(n-u') \neq 0$. If the inner product $\xi_q.(u - u')$ is not zero then (96) implies that

$$\Big|\varphi\big(2^{-j}\xi_1.(n-u)\big) - \varphi_q\big(2^{-j}\xi_q.(n-u')\big)\Big| > 0.$$

So if $x(u)$ and $x(u')$ are non-zero (95) implies that

$$\left|\varphi\big(2^{-j}\xi_1.(n-u)\big) - \varphi\big(2^{-j}\xi_q.(n-u')\big)\right| \geq \pi.$$

It follows from (96) that if $2^{-j}|u-u'| \leq 2\alpha$ then

$$2^{-j}\beta|\xi_q.(u-u')| \geq \pi,$$

which proves $|\xi_q.(u-u')| \geq C\,2^j$ for $C = \min(\pi\beta^{-1}, 2\alpha|\xi_q|)$, and hence Part (i).

Let us now prove property (ii). Since $\psi_q$ has a compact support it is included in $[-\gamma,\gamma]^\ell$ for $\gamma$ large enough. Since the support of $x$ are points of distance at least $\Delta$ it results that for any $n \in \mathbb{Z}^\ell$ and $2^j \leq \Delta\,\gamma^{-1}$, the product $x(u)\psi_{j,q}(n-u)$ is non-zero for at most one $u \in \mathbb{Z}^\ell$. It results that

$$\begin{aligned}\|x \star \psi_{j,q}\|_1 &= \sum_{n\in\Lambda_d}\Big|\sum_{u\in\Lambda_d} x(u)\,\psi_{j,q}(n-u)\Big| \\ &= \sum_{n\in\Lambda_d}\sum_{u\in\Lambda_d}|x(u)|\,|\psi_{j,q}(n-u)| = \|x\|_1\,\|\psi_{j,q}\|_1\,.\end{aligned}$$

The hypothesis (60) implies that $\|x'\|_1 = \|x' \star \psi_{j,q}\|_1$ for all $q \leq Q$ and $2^j \leq \Delta\,\min(1,\gamma^{-1})$. Applying Theorem 5.1 for $2^j \geq 2^{-1}\Delta\,\min(1,\gamma^{-1})$ proves that $x'(u)$ and $x'(u')$ are non-zero only for all $q \leq Q$ we have $\xi_q.(u-u') = 0$ or $|\xi_q.(u-u')| \geq C'\,\Delta$, where $C'$ does not depend upon $x$ and $x'$.

Since the $\{\xi_q\}_{q\leq Q}$ are $Q \geq \ell$ different rotations of a non-zero $\xi \in \mathbb{R}^\ell$, they define a frame of $\mathbb{R}^\ell$. It results that there exists $A$ and $B$ such that for any $v \in \mathbb{R}^\ell$

$$A\,|v| \leq \sum_{q\leq Q}|v.\xi_q| \leq B\,|v|\,. \tag{97}$$

This inequality applied to $v = u-u' \neq 0$ proves that there exists $q \leq Q$ such that $\xi_q.(u-u') \neq 0$. If $x(u) \neq 0$ and $x(u') \neq 0$ then we proved that if $\xi_q.(u-u') \neq 0$ then $|\xi_q.(u-u')| \geq C'\,\Delta$. The frame inequality (97) implies that $|u-u'| \geq B^{-1}\,C'\,\Delta$ which shows that any two points in the support of $x'$ have a distance at least $C\,\Delta$ with $C = C'B^{-1}$.

## References

[1] P.-A. Absil, R. Mahony, and B. Andrews, Convergence of the iterates of descent methods for analytic cost functions, *SIAM J. Optim.*, **16** (2005), no. 2, 531–547. Zbl 1092.90036 MR 2197994

[2] J. Andén and S. Mallat, Deep scattering spectrum, *IEEE Trans. Signal Process.*, **62** (2014), no. 16, 4114–4128. Zbl 1394.94040 MR 3260414

[3] L. Barbet, M. Dambrine, A. Daniilidis, and L. Rifford, Sard theorems for lipschitz functions and applications in optimization, *Israel J. Math.*, **212** (2016), no. 2, 757–790. Zbl 1353.58005 MR 3505402

[4] F. Barthe, O. Guédon, S. Mendelson, and A. Naor, A probabilistic approach to the geometry of $l_p^n$-ball, *Ann. Probab.*, **33** (2005), no. 2, 480–513. Zbl 1071.60010 MR 2123199

[5] G. Battle, *Wavelets and renormalization*, Series in Approximations and Decompositions, 10, World Scientific Publishing Co., Inc., River Edge, NJ, 1999. Zbl 0949.65145 MR 1688691

[6] M. Betancourt, A conceptual introduction to Hamiltonian Monte Carlo, 2017. arXiv:1701.02434

[7] E. Borel, Sur les principes de la theorie cinetique des gas, *Ann. Sci. École Norm. Sup. (3)*, **23** (1906), 9–33. Zbl 37.0944.01 MR 1509063

[8] P. Brémaud, L. Massoulié, and A. Ridolfi, Power spectra of random spike fields and related processes, *Adv. in Appl. Probab.*, **37** (2005), no. 4, 1116–1146. Zbl 1102.60030 MR 2193999

[9] J. Bruna and S. Mallat, Audio texture synthesis with scattering moments, 2013. arXiv:1311.0407

[10] J. Bruna and S. Mallat, Invariant scattering convolution networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **35** (2013), no. 8, 1872–1886.

[11] J. Bruna, S. Mallat, E. Bacry, and J.-F. Muzy, Intermittent process analysis with scattering moments, *Ann. Statist.*, **43** (2015), no. 1, 323–351. Zbl 1308.62168 MR 3311862

[12] J. V. Burke, A. S. Lewis, and M. L. Overton, A robust gradient sampling algorithm for nonsmooth, nonconvex optimization, *SIAM J. Optim.*, **15** (2005), no. 3, 751–779. Zbl 1078.65048 MR 2142859

[13] S. Chatterjee, A note about the uniform distribution on the intersection of a simplex and a sphere, *J. Topol. Anal.*, **9** (2017), no. 4, 717–738. Zbl 1379.35288 MR 3684622

[14] L. Chizat and F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport, in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), 3036–3046, Curran Associates, Inc., 2018.

[15] M. Creutz, Microcanonical Monte Carlo simulation, *Phys. Rev. Lett.*, **50** (1983), no. 19, 1411–1444. MR 701663

[16] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Jones and Bartlett Publishers, Boston, MA, 1993. Zbl 0793.60030 MR 1202429

[17] J. Deuschel, D. Stroock, and H. Zession, Microcanonical distributions for lattice gases, *Comm. Math. Phys.*, **139** (1991), no. 1, 83–101. Zbl 0727.60025 MR 1116411

[18] P. Diaconis and D. Freedman, A dozen de Finetti-style results in search of a theory, *Ann. Inst. H. Poincaré Probab. Statist.*, **23** (1987), no. 2, suppl., 397–423. Zbl 0619.60039 MR 898502

[19] M. Donsker and S. Varadhan, Large deviations for stationary Gaussian processes, *Comm. Math. Phys.*, **97** (1985), no. 1-2, 187–210. Zbl 0646.60030 MR 782966

[20] R. Ellis, *Entropy, large deviations, and statistical mechanics*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 271, Springer-Verlag, New York, 1985. Zbl 0566.60097 MR 793553

[21] B. Galerne, Y. Gousseau, and J.-M. Morel, Random phase textures: theory and synthesis, *IEEE Trans. Image Process.*, **20** (2011), no. 1, 257–267. Zbl 1372.94086 MR 2789729

[22] I. Gallagher, L. Saint-Raymond, and B. Texier, *From Newton to Boltzmann: hard spheres and short-range potentials*, Zürich Lectures in Advanced Mathematics, European Mathematical Society (EMS), Zürich, 2013. Zbl 1315.82001 MR 3157048

[23] L. Gatys, A. S. Ecker, and M. Bethge, Texture synthesis using convolutional neural networks, in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), 262–270, Curran Associates, Inc., 2015.

[24] S. Geman and D. Geman, Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.*, **6** (1984), no. 6, 721–741.

[25] H.-O. Georgii, *Gibbs measures and phase transitions*. Second edition, De Gruyter Studies in Mathematics, 9, Walter de Gruyter & Co., Berlin, 2011. Zbl 1225.60001 MR 2807681

[26] D. J. Heeger and J. R. Bergen, Pyramid-based texture analysis/synthesis, in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, 229–238, ACM Press, 1995.

[27] E. T. Jaynes, Information theory and statistical mechanics, *Phys. Rev. (2)*, **106** (1957), no. 4, 620–630. Zbl 0084.43701 MR 87305

[28] P. Kopietz, L. Bartosch, and F. Schütz, *Introduction to the functional renormalization group*, Lecture Notes in Physics, 798, Springer-Verlag, Berlin, 2010. Zbl 1196.82001 MR 2641839

[29] P. Kopietz, L. Bartosch, and F. Schütz, Mean-field theory and the gaussian approximation, in *Introduction to the functional renormalization group*, Lecture Notes in Physics, 798, Springer-Verlag, Berlin, 2010.

[30] O. E. Lanford III, Time evolution of large classical systems, in *Dynamical systems, theory and applications (Rencontres, Battelle Res. Inst., Seattle, Wash., 1974)*, 1–111, Lecture Notes in Phys., 38, Springer, Berlin, 1975. Zbl 0329.70011 MR 479206

[31] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, Gradient descent only converges to minimizers, in *Conference on Learning Theory*, 1246–1257, 2016.

[32] D. A. Levin and Y. Peres, *Markov chains and mixing times*. With contributions by Elizabeth L. Wilmer and a chapter on "Coupling from the past" by James G. Propp and David B. Wilson. Second edition, American Mathematical Society, Providence, RI, 2017. Zbl 1390.60001 MR 3726904

[33] S. Mallat, Group invariant scattering, *Comm. Pure Appl. Math.*, **65** (2012), no. 10, 1331–1398. Zbl 1282.47009 MR 2957703

[34] S. Mallat, S. Zhang, and G. Rochette, Phase harmonics and correlation invariants in convolutional neural networks, 2018. arXiv:1810.12136

[35] J. H. McDermott and E. P. Simoncelli, Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis, *Neuron*, **71** (2011), no. 5, 926–940.

[36] Y. Meyer, *Wavelets and operators*. Translated from the 1990 French original by D. H. Salinger, Cambridge Studies in Advanced Mathematics, 37, Cambridge University Press, Cambridge, 1992. Zbl 0776.42019 MR 1228209

[37] L. Onsager, Crystal statistics. I. A two-dimensional model with an order-disorder transition, *Phys. Rev. (2)*, **65** (1944), no. 3-4, 117–149. Zbl 0060.46001 MR 10315

[38] I. Panageas and G. Piliouras, Gradient descent converges to minimizers: The case of non-isolated critical points, 2016. arXiv:1605.00405

[39] J. Portilla and E. P. Simoncelli, A parametric texture model based on joint statistics of complex wavelet coefficients, *Int. J. Comput. Vis.*, **40** (2000), no. 1, 49–70. Zbl 1012.68698

[40] G. M. Rotskoff and E. Vanden-Eijnden, Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error, 2018. arXiv:1805.00915

[41] E. P. Simoncelli and B. A. Olshausen, Natural image statistics and neural representation, *Annual Review of Neuroscience*, **24** (2001), no. 1, 1193–1216.

[42] A. Sokol, *Advanced Probability*, Lecture Notes, 2013.

[43] D. W. Stroock and O. Zeitouni, Microcanonical distributions, Gibbs states, and the equivalence of ensembles, *Random walks, Brownian motion, and interacting particle systems*, 399–424, Progr. Probab., 28, Birkhäuser Boston, Boston, MA, 1991. Zbl 0733.00027 MR 1146461

[44] A. Tagliani, Hamburger moment problem and maximum entropy: On the existence conditions, *Appl. Math. Comput.*, **231** (2014), 111–116. Zbl 06892872 MR 3174016

[45] M. J. Wainwright and M. I. Jordan, Graphical models, exponential families, and variational inference, *Foundations and Trends® in Machine Learning*, **1** (2008), no. 1-2, 1–305.

[46] Y. Wang, W. Yin, and J. Zeng, Global convergence of ADMM in nonconvex nonsmooth optimization, *J. Sci. Comput.*, **78** (2019), no. 1, 29–63. MR 3902876

[47] M. Welling, Herding dynamical weights to learn, in *Proceedings of the 26th Annual International Conference on Machine Learning*, 1121–1128, ACM Press, 2009.

[48] S. Zhang and S. Mallat, Wavelet phase harmonic covariance models of stationary processes, 2019.

[49] S. C. Zhu, Y. Wu, and D. Mumford, Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling, *Int. J. Comput. Vis.*, **27** (1998), no. 2, 107–126.

Received 12 February, 2018; revised 11 January, 2019

J. Bruna, Courant Institute of Mathematical Sciences, New York University,
60 5th Avenue, New York, NY 10011, USA

E-mail: bruna@cims.nyu.edu

S. Mallat, Departement d'Informatique, Collège de France and Ecole Normale Supérieure,
45 rue d'Ulm, 75005 Paris, France

E-mail: stephane.mallat@ens.fr