Particle gradient descent model for point process generation

Antoine Brochard, Bartlomiej Blaszczyszyn, Stéphane Mallat, Sixin Zhang

Particle gradient descent model for point process generation

Antoine Brochard *
Bartłomiej Błaszczyszyn †
Stéphane Mallat ‡
Sixin Zhang §

October 27, 2020

Abstract

This paper introduces a generative model for planar point processes in a square window, built upon a single realization of a stationary, ergodic point process observed in this window. Inspired by recent advances in gradient descent methods for maximum entropy models, we propose a method to generate similar point patterns by jointly moving particles of an initial Poisson configuration towards a target counting measure. The target measure is generated via a deterministic gradient descent algorithm, so as to match a set of statistics of the given, observed realization. Our statistics are estimators of the multi-scale wavelet phase harmonic covariance, recently proposed in image modeling. They allow one to capture geometric structures through multi-scale interactions between wavelet coefficients. Both our statistics and the gradient descent algorithm scale better with the number of observed points than the classical k-nearest neighbour distances previously used in generative models for point processes, based on the rejection sampling or simulated-annealing. The overall quality of our model is evaluated on point processes with various geometric structures through spectral and topological data analysis.

1 Introduction

In order to perform statistical analysis of a random phenomenon of which we only have a single observation, it is often useful to build a probabilistic model from which we can simulate approximations of the underlying process. This problem has been previously studied in a wide range of domains [1, 5, 22, 32, 35]. A classical method in the point process literature is based on transforming an initial random counting measure by successively replacing a randomly selected atom of the measure by a new, randomly positioned atom, so as to match some statistical descriptors of the observed sample [33]. This approach works well on relatively small-scale and simple geometrical structures, however, when facing complex multiscale geometries formed by the particles in turbulence flows [15, 16, 25, 27] and cosmology [30, 31], we need to develop a more efficient and accurate model. We shall study this problem in the mathematical framework of the classical maximum-entropy models [21], and propose a particle gradient-descent model, which is adapted from recent progresses in modeling non-Gaussian ergodic and stationary processes [10, 36].

Suggested by statistical physics, maximum-entropy models are distributions of maximal *entropy* that match a set of prescribed statistical descriptors. Intuitively, this means that the model is 'as random as possible' under certain constraints, in the sense that there are no constraints other than the ones based on the information captured by the descriptors. Two standard models are the *macrocanonical model*, with expectation constraints, and the *microcanonical model*, with pathwise

* *Huawei R&D France* and *Inria/ENS*, Paris, France
† *Inria/ENS, Paris, France*
‡ *Collège de France* and *ENS* Paris, France
§ *IRIT*, Université de Toulouse, CNRS, Toulouse, France

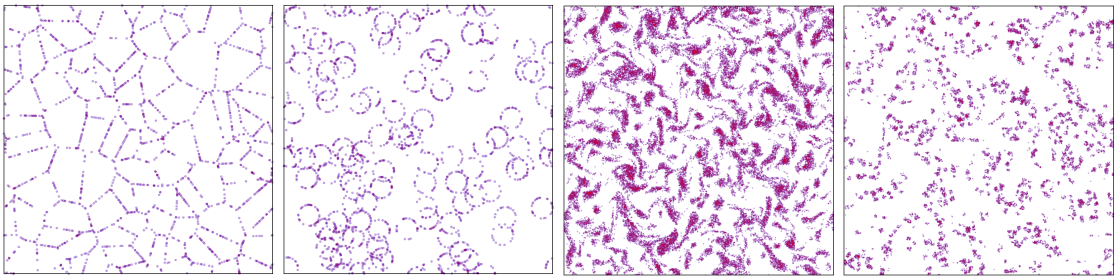

Figure 1: Samples of point processes of various geometries. The number of points range from 1000-40000 (see a summary in Table 1 in Section 6).

constraints. In the context of point processes, where Shannon's entropy is not well defined, it is customary to replace it by Kullback-Leibler (KL) divergence with respect to the reference Poisson point process distribution, and minimize this latter under the macro- or microcanonical constraints. In this case, the solutions to both models are (two different) Gibbs point processes (having density with respect to the reference Poisson distribution). Although mathematically elegant, these Gibbs point processes are hard to sample from. Variants of rejection sampling, simulated annealing and MCMC algorithms are suggested in this regard [14], all being extremely time consuming if the number of points grows large.

In [10], the authors study a *gradient descent model* as an approximation of the microcanonical model for ergodic and stationary processes. For processes on the two-dimensional square lattice (modeling pixel images), it consists in transporting an initial high-entropy probability measure such as Gaussian white noise through gradient descent on the amplitude of each pixel. Gradient descent models are very popular in image and time-series modeling ([5,18,24,28,36]) because they allow for fast sampling of the model, as opposed to MCMC algorithms. While this approach has been already applied in [10] to point processes on the plane subjected to the lattice locations, it is not adapted to the nature of point processes, for which the matter of interest is the positions of the points in the continuous plane. Thus, optimizing the amplitude of each pixel does not take advantage of the a priori information that the process is an atomic measure.

The main contribution of this paper is to propose an approach taking advantage of the efficiency of gradient descent algorithms, while preserving the continuum nature of point processes. We call it *particle gradient descent*. Similar to the work [33], all the particles (points) of an initial realization (of the Poisson point process) are moved continuously on the plane. However, as opposed to using acceptance-rejection sampling computed in a sequential order, we propose a deterministic and parallel gradient descent algorithm to make the sampling more efficient. Moreover, we adapt the descriptors developed from multi-scale wavelet analysis to capture complex geometric structures formed by a large number of particles.

The choice of the descriptors is essential to generate high-quality samples. Common statistics, more related to point process theory, such as the nearest neighbour distance distribution, or empty space distribution, could be used. However, if the number of points in the configuration becomes large, these statistics will only capture small scale geometric information. In [33], the authors advocate the use of k-nearest neighbours, which could become intractable for a (necessarily) large k. Instead, we use the multi-scale wavelet phase harmonic descriptors, which were recently introduced in [24,36] . To match these descriptors for point processes, we propose a complete numerical scheme allowing one to perform efficiently the gradient descent of these descriptors with respect to the positions of all particles. It is based on a differentiable discretization of atomic measures, combined with multiscale optimization in the gradient descent, meant to avoid undesirable shallow minima.

We evaluate our model on some distributions exhibiting complex geometric patterns, like Cox point processes on the edges of Poisson-Voronoi tessellations and on the Boolean model with circular grains, as well as on Mattern hard-core and Mattern cluster processes driven by Poisson processes with turbulent intensities (see Figure 1 for a few examples). Besides the visual inspection of the samples from our generative model, which is an important and standard analysis to assess the

closeness of the model distribution to that of the observation, we evaluate second order correlations by estimating the Bartlett spectrum for the original samples and the generated approximations. In order to study in more details geometric structures, we compare the persistent homology diagrams, as it has been proven useful for topological data analysis [11].

The remaining part of the paper is organized as follows: in Section 2 we recall some basic notions from the theory of point processes and reformalize the macro and microcanonical models in this setting . Section 3 presents our generative model. We discuss modeling issues related to the choice of the descriptor and sampling algorithm, and then state an invariance conservation property, which extends the result in [10] to the continuous domain. Based on [36], in Section 4 we present our descriptor to characterize geometric structures of point processes. Section 5 describes a complete numerical scheme allowing one to perform efficiently the gradient descent of these descriptors. In Section 6 we evaluate our model numerically. We complete the paper with three sections in the Appendix: Appendix A details a regularised variant of our gradient descent model. Appendix B gives a proof of the invariance result of Section 3. Finally Appendix C recalls some basic notions of Fourier theory for point processes, and presents a brief comparison of spectral and topological analysis of the data.

2 Point process framework and notations

2.1 General definitions

In this section we define the elementary objects of point process theory and the notations that we will use in this paper. A more detailed introduction to stochastic geometry and point processes can be found in [4,13]. In this paper, we focus on 2d point processes. The methodology we propose can be readily extended to 1d or 3d.

Configurations of points (on the plane) are represented as counting measures on $(\mathbb{R}^2, \mathcal{B})$, with $\mathcal{B}$ denoting the natural Borel σ-algebra on $\mathbb{R}^2$. Recall that counting measures are locally finite measures taking values in $\bar{\mathbb{N}} := \mathbb{N} \cup \{+\infty\}$. Let $\mathbb{M}$ denote the space of all such measures on $(\mathbb{R}^2, \mathcal{B})$, endowed with the σ-algebra $\mathcal{M}$ generated by the mappings $\mu \mapsto \mu(B)$, for $B \in \mathcal{B}$. For $\mu \in \mathbb{M}$, we will often use the following representation:

$$\mu = \sum_{1 \leq i \leq I} \delta_{x_i},\ I \in \bar{\mathbb{N}}, \tag{1}$$

where δ_x is the Dirac measure having a unit atom at x. Recall, a *push-forward* $F_{\#}\mu$ of a point measure μ by a (measurable) function $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is simply the displacement of its atoms by the function F

$$F_{\#}\mu = \sum_i \delta_{F(x_i)}.$$

A counting measure $\mu \in \mathbb{M}$ is called simple if for all $x \in \mathbb{R}^2$ $\mu(\{x\}) = 0$ or 1 (in other words all atoms of μ in the representation (1) are distinct). Simple counting measures can be identified with their supports $\mathrm{Supp}(\mu) := \{x \in \mathbb{R}^2 : \mu(\{x\}) > 0\}$ and in this regard we shall also write $x \in \mu$ if x is an atom of μ, i.e., if $\mu(\{x\}) > 0$.

For $\mu \in \mathbb{M}$ and $x \in \mathbb{R}^2$, we define the translation $S_x\mu$ of μ by x, i.e. $S_x\mu(B) := \mu(B + x)$.

A point process Φ is a measurable mapping from an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{M}, \mathcal{M})$. We will denote by $\mathcal{L}_\Phi$ the distribution of Φ, that is the pushforward of the probability measure $\mathbb{P}$ by Φ on $(\mathbb{M}, \mathcal{M})$. We shall denote by $\mathcal{D}$ the space of probability distributions on $(\mathbb{M}, \mathcal{M})$. We say that a point process Φ is simple if $\mathbb{P}(\Phi \text{ is a simple measure}) = 1$. Point process Φ is termed *stationary* if its distribution $\mathcal{L}_\Phi$ is invariant with respect to all shifts S_x, $x \in \mathbb{R}^2$. It is termed *ergodic* if the empirical averages (of real, measurable functions f on $\mathbb{M}$, integrable with respect to $\mathcal{L}_\Phi$) over windows $W_s = [-s, s[\times[-s, s[$ increasing to $\mathbb{R}^2$ converge almost surely to the mathematical expectations

$$\lim_{s \to \infty} \frac{1}{|W_s|} \int_{W_s} f(S_x\Phi)\, dx = \mathbb{E}[f(\Phi)] = \int_{\mathbb{M}} f(\mu)\, \mathcal{L}_\Phi(d\mu), \tag{2}$$

where $|W_s|$ stands for the Lebesgue measure of W_s, see [4, Chapter 8] fore more details.

For a given $s > 0$, we denote by $\mathbb{M}^s$ the set of counting measures on W_s, and $\mathcal{M}^s$ its induced σ-algebra. Unless otherwise stated, we will consider W_s with addition and scalar multiplication modulo W_s. Also we shall denote by $\bar{S}_x$ the corresponding shift operator on $\mathbb{M}^s$.

For any integer $n \geq 1$ and any $z \in \mathbb{C}^n$, we note $|z|$ the Euclidean norm of z.

Let Φ be a point process on $\mathbb{R}^2$. One can only observe realizations of Φ on bounded subsets of $\mathbb{R}^2$. For the remaining of this paper, we shall consider realizations of point processes observed on a finite square window W_s, for some $s > 0$. We shall note $\bar{\Phi}$ the restriction of Φ to W_s, that is $\bar{\Phi}$ is a point process on W_s such that, $\forall\, n > 0, \forall\, (B_1, ..., B_n) \in \mathcal{B}(W_s)^n, (\Phi(B_1), ..., \Phi(B_n)) = (\bar{\Phi}(B_1), ..., \bar{\Phi}(B_n))$ in distribution (where $\mathcal{B}(W_s)$ stands for the Borel σ-algebra on W_s). A realization of Φ observed on W_s is therefore a realization of $\bar{\Phi}$, and will be noted $\bar{\phi}$.

2.2 Notes on maximum entropy models for point processes

The notion of entropy is naturally defined only for random objects in discrete state spaces. Even if a mixture of the differential and discrete entropy can be considered for point processes [3], it is more natural to consider in this context the Kullback-Leibler (KL) divergence with respect to a reference distribution. naturally taken to be the homogeneous Poisson point process distribution ([14]). More specifically, let us denote by $\mathcal{L}_0$ the Poisson distribution on W_s. We define the KL divergence of a point process $\bar{\Phi}$ on W_s with distribution $\mathcal{L}_\Phi$,

$$\mathrm{KL}(\mathcal{L}_\Phi; \mathcal{L}_0) := \int_{\mathbb{M}^s} \frac{d\mathcal{L}_\Phi}{d\mathcal{L}_0}(\mu) \log \frac{d\mathcal{L}_\Phi}{d\mathcal{L}_0}(\mu)\, \mathcal{L}_0(d\mu); \tag{3}$$

provided $\mathcal{L}_\Phi$ is absolutely continuous w.r.t. $\mathcal{L}_0$, denoting by $\frac{d\mathcal{L}_\Phi}{d\mathcal{L}_0}$ the corresponding density. (Otherwise KL is set to ∞.)

Remark 2.1. *In this paper, we shall consider that the number of points of our model in W_s is fixed. In such case it is customary to take homogeneous Poisson point process distribution conditioned to have exactly n points in W_s as the reference measure, which is equivalent to n points sampled uniformly, independently in W_s. We will note this distribution $\mathcal{L}_0^n$.*

With the KL divergence as a notion of entropy for point processes, we can now define the macrocanonical and microcanonical models. These models are distributions of maximum entropy under different types of constraints. When considering these distributions as models for a point process $\bar{\Phi}$, the constraints are usually built as functions of the distribution of $\bar{\Phi}$, or functions of samples from $\bar{\Phi}$. Consider a mapping $K : \mathbb{M}^s \longrightarrow \mathbb{C}^d$, for some $d < \infty$, the macro- and microcanonical models are defined as follows:

Macrocanonical model In this model one is looking for a point process Ξ with distribution $\mathcal{L}$ on $\mathbb{M}^s$ that *minimizes* the KL divergence $\mathrm{KL}(\mathcal{L}, \mathcal{L}_0)$ under *average constraints*:

$$\arg\min_{\mathcal{L}} \quad \mathrm{KL}(\mathcal{L}, \mathcal{L}_0) \tag{4}$$

$$\text{given} \quad \mathbb{E}(K(\Xi)) = a, \tag{5}$$

for some vector of constraints $a = \mathbb{E}(K(\bar{\Phi}))$. The solution of this problem is the Gibbs (point process) distribution[1] $\mathcal{L}_G$ on $\mathbb{M}^s$ having the density $e^{U(\mu)}/Z$ with respect to $\mathcal{L}_0$, i.e. such that

$$\int_{\mathbb{M}^s} f(\mu)\, \mathcal{L}_G(d\mu) = 1/Z \int_{\mathbb{M}^s} f(\mu) e^{U(\mu)}\, \mathcal{L}_0(d\mu),$$

with $U(\mu) = \lambda^*(K(\mu) - a)$, where the vector λ^* is given by

$$\lambda^* = \arg\min_{\lambda} \log \int_{\mathbb{M}^s} e^{\lambda(K(\mu)-a)}\, \mathcal{L}_0(d\mu)$$

and $Z = \int_{\mathbb{M}^s} e^{U(\mu)}\, \mathcal{L}_0(d\mu)$. When given only one realization $\bar{\phi}$ of Φ, we take as the vector of constraints $a = K(\bar{\phi})$. Building an accurate model therefore requires that $K(\bar{\phi})$ concentrates around $\mathbb{E}(K(\bar{\Phi}))$.

[1] Conditions for the existence of the solution of this form is left for future work. This problem is related to existing works on Gibbs point processes [14].

Microcanonical model We define the microcanonical set of level ϵ, for some $\epsilon > 0$ and observation $\bar{\phi}$ such that $a = K(\bar{\phi})$:

$$\Omega_\epsilon := \{\mu \in \mathbb{M}_s \,:\, |K(\mu) - a| \leq \epsilon\}. \tag{6}$$

In the microcanonical model one is looking for a distribution $\mathcal{L}$ that minimizes the KL divergence with respect to the reference distribution $\mathcal{L}_0$ under a *pathwise constraint* requiring $\mathcal{L}$ to be constrained to Ω_ϵ:

$$\arg\min_{\mathcal{L}} \quad \mathrm{KL}(\mathcal{L}, \mathcal{L}_0) \tag{7}$$

$$\text{given} \quad \int_{\mathbb{M}^s} \mathbb{1}(\mu \in \Omega_\epsilon)\mathcal{L}(d\mu) = 1, \tag{8}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Here, the solution is again a Gibbs point process distribution, with the density $V(\mu)/Z$ with respect to $\mathcal{L}_0$ which being simply the conditioning of $\mathcal{L}_0$ to Ω_ϵ: $V(\mu) = \mathbb{1}(\mu \in \Omega_\epsilon)$ and $Z = \mathcal{L}_0(\Omega_\epsilon)$. [2] If K is defined as spatial averages of integrable functions (defined in (10)), and if additionally Φ is ergodic and s is large enough such that $K(\bar{\phi}) \simeq \mathbb{E}(K(\Phi))$ $K(\bar{\phi}) \simeq \mathbb{E}(K(\bar{\Phi}))$, then the microcanonical model can be considered as a good approximation of the macrocanonical model, under the Boltzmann equivalence conjecture. We refer the reader to [10] for more details on this subject.

Sampling from both macro- and microcanonical models is a challenging task. For this reason, in what follows we consider a different approach based on a Poisson point process transport via a gradient descent algorithm. It has been argued in [10] that this model may be seen as an approximation of the microcanonical model.

3 Generative model for point processes

Let Φ be an ergodic stationary point process on $\mathbb{R}^2$ with unknown distribution, and $\bar{\phi} \in \mathbb{M}^s$ be a realization of Φ observed on a finite window W_s. $\bar{\phi}$ is therefore a realization of a point process $\bar{\Phi}$ on W_s, defined as the restriction of Φ to W_s.

We want to build a generative model for $\bar{\Phi}$, based on our single sample $\bar{\phi}$, so that we can generate approximate samples from $\bar{\Phi}$. To that end, we shall compute some descriptor K of $\bar{\phi}$ and build a gradient descent model that will allow us to transport any initial random configuration of points "towards" $\bar{\phi}$ by minimizing the distance on the space of the descriptor. Since we want to view this model as an approximation of the microcanonical model (i.e. sample from Ω_ϵ while minimizing the KL divergence w.r.t. $\mathcal{L}_0^{\bar{\phi}(W_s)}$, see Section 2.2) it is natural to choose as the initial measure $\mathcal{L}_0^{\bar{\phi}(W_s)}$. This is in analogy with taking initial distribution with the highest possible entropy, i.e., Gaussian white noise, in the lattice model considered in [10]. Note however, that the gradient descent model is not in general the microcanonical maximum entropy model.

More specifically, from the mapping $K : \mathbb{M}^s \longrightarrow \mathbb{C}^d$ for the given configuration of points $\bar{\phi} \in \mathbb{M}^s$, let us define a mapping from $\mathbb{M}^s$ to $[0, \infty)$

$$E_{\bar{\phi}}(\mu) := \frac{1}{2}|K(\mu) - K(\bar{\phi})|^2. \tag{9}$$

We can interpret the function $E_{\bar{\phi}}$ as the energy on $\mathbb{M}^s$ expressing the "similarity" of any $\mu \in \mathbb{M}^s$ with respect to the given configuration $\bar{\phi}$.

Our generative model consists in sampling points of the initial configuration $\bar{\phi}_0$ independently, uniformly in W_s, their number being equal to this of $\bar{\phi}$, and calculating a sequence of push-forward point measures $\bar{\phi}_{n+1} = F^n_\# \bar{\phi}_n$, $n \geq 0$, by some function $F^n : W_s \longrightarrow W_s$ (which depend on $\bar{\phi}$ and $\bar{\phi}_n$), such that $E_{\bar{\phi}}(\bar{\phi}_n)$ converges to a (typically local) minimum. The function F^n corresponds to the one-step gradient descent of $E_{\bar{\phi}}(\bar{\phi}_{n-1})$ with respect to the locations of points of $\bar{\phi}_n$, as will be explained Section 3.1.

In principle, the sequence of the above point processes can be considered for any descriptor K. However, in order to be able to consider $\bar{\phi}_n$, for large $n \geq 0$, as an approximate sampling from

[2] Note (7) can be reduced to (4) with $K(\mu) = 1(\mu \in \Omega_\epsilon)$ and $a = 1$.

the unknown distribution of $\bar{\Phi}$, we make the following, more or less formalized postulates, that we relate to the classical macro- and microcanonical models assumptions:

(P1) *Concentration property*: The value of $K(\bar{\Phi})$ should concentrate around its mean, i.e. $K(\bar{\Phi}) \simeq \mathbb{E}[K(\bar{\Phi})]$ with high probability. A natural assumption is that the variance of $K(\bar{\Phi})$ is small.

(P2) *Sufficiency property*: The moment of the descriptor, $\mathbb{E}(K(\bar{\Phi}))$, should *characterize the unknown distribution* as completely as possible.[3] It requires that K has a strong (distributional) discriminate power. (In the marcroanonical model, it means that we want $KL(\mathcal{L}_G, \mathcal{L}_\Phi) \simeq 0$, cf. Paragraph 2.2.)

In addition to the above two postulates (P1), (P2) regarding explicitly the descriptor K, we have to assume the following property of the gradient descent algorithm (which implicitly involves the descriptor K too):

(P3) The loss of the gradient descent model (i.e., the value $E_{\bar{\phi}}(\bar{\phi}_n)$) is small enough for large n with high enough probability (or frequency) regarding the initial configuration. This assumption should be understood as making our generative model explore realizations of Ω_ϵ.

Given (P3), we do not know if our gradient descent algorithm "explores Ω_ϵ uniformly" as postulated in the microcanonical model. This is our idealized postulate.

A natural framework allowing one to address (P1) and (P2) is this of $K = (K_1, \dots, K_d)$ being a vector of empirical averages

$$K_i(\mu) = \frac{1}{|W_s|} \int_{W_s} f_i(\bar{S}_x \mu)\, dx \qquad \mu \in \mathbb{M}^s, \tag{10}$$

for a sufficiently rich class of functions f_i on $\mathbb{M}^s$, and relying on the ergodic assumption (2) regarding Φ. [4]

These properties are necessary in order to have a model that reproduces typical geometric structures in Φ, and generates diverse samples (i.e. that are distinct from one another). We shall rather present in Section 4 a specific descriptor meant to satisfy these postulates, and statistically verify in Section 6 the quality of our generative model in relation to the postulates.

3.1 Particle gradient descent model

Recall that in this section and in what follows, W_s is interpreted as endowed with the addition and scalar multiplication modulo W_s. Moreover, $\bar{S}_x$ deonotes the corresponding periodic shift by the vector x on $\mathbb{M}^s$.

For $\mu \in \mathbb{M}^s$ and any $x \in \mathrm{Supp}(\mu)$, we define the following functions:

$$\begin{array}{rccc} h_x^\mu & : & \mathbb{R}^2 & \longrightarrow & \mathbb{M}^s \\ & & y & \longmapsto & \mu - \delta_x + \delta_{x+y}, \end{array}$$

$$\begin{array}{rccc} K_x^\mu & : & \mathbb{R}^2 & \longrightarrow & \mathbb{C}^d \equiv \mathbb{R}^{2d} \\ & & y & \longmapsto & K \circ h_x^\mu(y), \end{array}$$

$$\begin{array}{rccc} E_x^\mu & : & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^+ \\ & & y & \longmapsto & E_{\bar{\phi}} \circ h_x^\mu(y). \end{array}$$

The function K_x^μ can be complex valued. However, as our energy function is the square Euclidean norm, it is equivalent to consider that K_x^μ has values in $\mathbb{R}^{2d}$. Moreover, we assume

[3] Giving a clear mathematical meaning to the statement 'as completely as possible' is possible, but it is beyond the scope of this paper.

[4] Note the boundary problem with respect to the ergodic result (2) due to the fact that we do not observe points outside W_s. Different boundary conditions can be assumed: extending $\mu \in \mathbb{M}^s$ by setting it null outside W_s or using the periodic extension (which corresponds to replacing S_x by the periodic shift $\bar{S}_x$ in (10)). The latter option will be our default one. In any case, a proper justification of the corresponding LLN with some specific boundary condition usually requires stronger, mixing properties for the underlying point process distribution.

in what follows that the function K is such that for all $\mu \in \mathbb{M}^s$ and all $x \in \mathrm{Supp}(\mu)$, K_x^μ is differentiable. We can then define, for any $\mu \in \mathbb{M}^s$ and any $x \in W_s$

$$\nabla_x K(\mu) := \begin{cases} Jac[K_x^\mu](0) & \text{if } x \in \mathrm{Supp}(\mu) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

and

$$\nabla_x E_{\bar\phi}(\mu) := \begin{cases} Jac[E_x^\mu](0) = (\nabla_x K(\mu))^t (K(\mu) - K(\bar\phi)) & \text{if } x \in \mathrm{Supp}(\mu) \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where $Jac[f]$ denotes the Jacobian matrix of the function f.

Finally, define the mapping

$$\begin{array}{rccc} F & : & \mathbb{M}^s & \longrightarrow & \mathbb{M}^s \\ & & \mu = \sum_i \delta_{x_i} & \longmapsto & \displaystyle\sum_i \delta_{x_i - \gamma \nabla_{x_i} E_{\bar\phi}(\mu)} \end{array}$$

for some gradient step $\gamma > 0$.

Remark 3.1 (Pushforwad of the point measures)**.** *The measure $F(\mu)$ can be seen as the push-forward $F_{\mu\#}\mu$ of the measure μ by the mapping*

$$\begin{array}{rccc} F_\mu & : & W_s & \longrightarrow & W_s \\ & & x & \longmapsto & x - \gamma \nabla_x E_{\bar\phi}(\mu). \end{array}$$

Note that the function F_μ depends on the measure μ which is pushed forward.

Remark 3.2 (Pushforward of the point process distributions)**.** *The pushforward operation $F_{\#\mu}$ on $\mathbb{M}^s$ induces the corresponding pushforward operation on the probability measures on $\mathbb{M}^s$, which are distributions of point processes. We denote this latter by $F_{\#\#}$: For a probability law $\mathcal{L}$ on $\mathbb{M}^s$ $F_{\#\#}\mathcal{L}(\Gamma) := \mathcal{L}(\{\mu \in \mathbb{M}^s : F_{\mu\#}\mu \in \Gamma\})$ for $\Gamma \in \mathcal{M}^s$.*

For any initial point measure $\bar\phi_0 \in \mathbb{M}^s$ we define the successive point measures:

$$\bar\phi_n := F_{\bar\phi_{n-1}\#}\bar\phi_{n-1}, \qquad n \geq 1, \tag{13}$$

Similarly, for a probability law $\mathcal{L}_0$ on $\mathbb{M}^s$ we define the successive probability laws

$$\mathcal{L}_n := F_{\#\#}\mathcal{L}_{n-1}, \qquad n =\geq 1 \tag{14}$$

Note $\mathcal{L}_n = \mathcal{L}_{\Phi_n}$ is the distribution of the point process $\bar\Phi_n$ obtained by n iterations of (13) starting from $\bar\Phi_0$ having law $\mathcal{L}_0 = \mathcal{L}_{\Phi_0}$.

Our implicit assumption is that for almost all $\bar\phi_0$ from distribution $\bar\Phi_0$ the sequence $\bar\phi_n$ converges to some $\bar\phi_\infty$ satisfying $E_{\bar\phi}(\bar\phi_\infty) \approx 0$; cf. postulate (P3).

3.2 Leveraging invariances

One can leverage some a priori known invariance properties of Φ (for instance stationarity or isotropy), by building a descriptor K that satisfy the same invariance properties than Φ.

This requires some explanation, since invariance properties of the distribution of Φ do not, in general, imply any natural invariance of its truncation $\bar\Phi$ to W_s. Indeed, while some invariances can be observed on the torus for the distribution of Φ on $\mathbb{R}^2$ (the most popular being translation invariance), it does not imply the same for $\bar\Phi$ with respect to the translation on W_s. The latter, called in this paper *circular stationarity*, requires also Φ to be periodic. However, circular stationarity of the generated point process on large window W_s (as a distributional approximation of $\bar\Phi$) can be considered as a desirable ersatz of the stationarity of Φ. It can be proved under some natural assumptions regarding the descriptor K and the distribution of the initial configuration Φ_0 of our generative model. Indeed, in what follows we shall formulate a result saying that, when K and

the distribution of Φ_0 are invariant with respect to some subset of *rigid circular transformations* on W_s, then the resulting model satisfies this property as well.

More specifically, a *rigid circluar transformation* on W_s is an invertible operator T on W_s of the form $Tx := Ax + x_0$ for some orthogonal matrix A with entries in $\{-1, 0, 1\}$[5], and $x_0 \in W_s$ (recall that additions and multiplications are modulo W_s here). We say that:

- The initial probability law $\mathcal{L}_{\bar{\Phi}_0}$ of the model is invariant to the action of T if $\forall\, \Gamma \in \mathbb{M}^s$, $\mathcal{L}_{\Phi_0}(T_{\#}^{-1}(\Gamma)) = \mathcal{L}_{\Phi_0}(\Gamma)$. [6]
- The descriptor K is invariant to the action of T if $\forall\, \mu \in \mathbb{M}^s$, $K(T_{\#}\mu) = K(\mu)$.

Theorem 3.3. *Let T be a rigid circular transformation. Let $\bar{\Phi}_0$ be a point process on W_s such that its distribution $\mathcal{L}_{\bar{\Phi}_0}$ is invariant to the action of T and let K be a descriptor defined on $\mathbb{M}^s$ invariant to the action of T. Then, for all $n \in \mathbb{N}$, $\mathcal{L}_{\bar{\Phi}_n}$ defined as the push-forward of $\mathcal{L}_{\bar{\Phi}_0}$ by* (14) *is invariant to the action of T.*

A proof of the above result is given in Appendix B . [7] Observe, the invariance of the distribution of the point process $\bar{\Phi}_n$ increases the diversity of the generative model samples. Our descriptor K proposed in Section 4.2 will be invariant with respect to all circular translations. This will be achieved by computing statistics of $\bar{\Phi}$ in the form of spatial averages (10) with periodic boundary condition. A drawback is that such a boundary condition introduces some bias to the spatial average (10) as an estimator of $\mathbb{E}[K(\bar{\Phi})]$ in the case of a non periodic Φ. One can expect, however, that when the window size is large enough and spatial correlation of the pattern are not too large, this border effect becomes negligible. If it is not the case, other boundary corrections have to be applied. For more a more detailed discussion on this issue, see the paragraph Non-periodic integration on page 12.

4 Wavelet phase harmonic descriptors

In this section we present a family of descriptors that we will use, in conjunction with the particle gradient descent model, to capture and reproduce complex geometries of point processes.

Classical descriptors for spatial point process usually include statistics more or less directly related to pair correlation function, such as the Ripley's K or L function, or the radial distribution function; see [12, Section 4.5]. All of these functions only capture second order correlations of the process. Other usual functions are the empty space function or the k-nearest neighbors function; see [12, Section 2.3.4 and 4.1.7]. In [33], the authors advocate the use of the k-nearest neighbors distance function, with a k significantly larger than 1. It is clear that this function becomes informative regarding higher order correlations of points as k grows. Ultimately it characterises the pairwise distance matrix, and thus the point pattern, up to isometries. However, the spatial range of correlations that this function captures depends on the number of points in the observation sample, and as it becomes large, it would require a large k to capture enough information. The number of descriptors would therefore become too large, leading to high computation time and high variance in the estimators.

For this reason, we choose in this paper to use descriptors for which the spatial range of structure captured is independent of the intensity of the process, and the computational time is linear in the number of points. As a result, this method would become much faster for large samples, as the number of statistics would remain constant. These descriptors, built upon wavelet transform of a random configuration, are adapted from [36]. They have shown high quality results in a number of problems including turbulence modelling.

We begin, in Section 4.1, by presenting wavelet transform for counting measures, and their so called phase harmonics, which are derived from wavelet transforms by applying a multiplication

[5] These conditions are necessary in order for T to be a well defined invertible operator. It encapsulates translations, flips, and orthogonal rotations.

[6] Note, $T_{\#}^{-1}(\Gamma) = \{\mu \in \mathbb{M}^s : T_{\#}\mu \in \Gamma\}$. For an invertible T (for which T^{-1} exists) and any $\mu \in \mathbb{M}$, we shall also write $T_{\#}^{-1}\mu := T_{\#}^{-1}\{\mu\} = (T^{-1})_{\#}\mu$.

[7] The result itself is inspired by Theorem 3.4 from [36], where the preservation of invariance is proven for the gradient descent model in the pixel domain.

operator on their phase. In Section 4.2, we explain how wavelet phase harmonics can be used to capture dependencies between the wavelet coefficients of the counting measures, and detail the choice of the descriptors that we use for numerical experiments. We propose in Section 4.2.2 several variants of the model to achieve a good balance between the concentration (P1) and the sufficiency (P2) of the descriptors.

4.1 Wavelet transforms and their phase harmonics

We review the wavelet transform [23, 26] for counting measures and then review the wavelet phase harmonics [24]. They allow us to define in the next section the descriptors that we propose to use in conjunction with our generative model described in Section 3.1

4.1.1 Wavelet transform

The wavelet transform is a powerful tool in image processing to analyze signals presenting local geometric structures of different scales. To capture edge-like structures in a planar point process, we shall use bump steerable wavelets introduced in [24]. They are defined by the translations, dilations and rotations of a complex analytic function $\psi(x) \in \mathbb{C}$ with $\int \psi(x)\, dx = 0$.

In what follows we first define the wavelet transform for the counting measures in $\mathbb{M}$ and then in $\mathbb{M}^s$. Let us denote the Fourier transform of ψ for $\omega \in \mathbb{R}^2$ by $\widehat{\psi}(\omega) = \int \psi(x) e^{-i\langle \omega, x\rangle} dx$, where $\langle a, b\rangle$ denotes the Euclidean inner product between two vectors $a \in \mathbb{R}^2$ and $b \in \mathbb{R}^2$. The function ψ is centered at a frequency $\xi_0 \in \mathbb{R}^2$, and it has a compact support in the frequency domain, as well as a fast spacial decay. Assume that $|\psi(x)|$ is negligible if $|x| > C$, for some $C > 0$.

Let r_θ denote the rotation by angle θ in $\mathbb{R}^2$. Multiscale steerable wavelets are derived from ψ with dilations by factors 2^j for $j \in \mathbb{Z}$, and rotations r_θ over angles $\theta = 2\ell\pi/L$ for $0 \leq \ell < L$, where L is the number of angles between $[0, 2\pi)$. The wavelet at scale j and angle θ is indexed by its central frequency $\lambda := 2^{-j} r_{-\theta}\, \xi_0 \in \mathbb{R}^2$, and it is defined by

$$\psi_\lambda(x) = 2^{-2j} \psi(2^{-j} r_\theta x) \quad \Rightarrow \quad \widehat{\psi}_\lambda(\omega) = \widehat{\psi}(2^j r_\theta \omega).$$

Since $\widehat{\psi}(\omega)$ is centered around ξ_0, it results that $\widehat{\psi}_\lambda(\omega)$ is centered around the frequency λ. The wavelet ψ_λ at scale j is self-similar to ψ in the spacial domain and its amplitude is thus negligible for $|x| > 2^j C$.

For a counting measure $\bar{\phi}$ observed in W_s, we typically consider only the wavelets having spatial support (more precisely, where the wavelet norm is non negligible) contained in W_s by limiting the scale $j < J$ such that $2^J C \leq 2s$. Scales equal or larger than J are carried by a low-pass filter whose frequency support is centered at $\lambda = 0$. It is denoted by ψ_0. Let Λ be a frequency-space index set including $\lambda = 2^{-j} r_{-\theta} \xi_0$ for $0 \leq j < J$, $0 \leq \ell < L$, and $\lambda = 0$. As we eliminate $j < 0$ in Λ to ignore structures smaller than C in W_s, the parameter ξ_0 will be adjusted in Section 5.1 for a suitable choice of C.

The *wavelet transform* of a counting measure $\phi \in \mathbb{M}$ is a family of functions obtained by the convolution of ϕ with $\{\psi_\lambda\}_{\lambda \in \Lambda}$, i.e.

$$\phi \star \psi_\lambda(x) = \int_{\mathbb{R}^2} \psi_\lambda(x - y) \phi(dy), \quad \lambda \in \Lambda. \tag{15}$$

This integral can be interpreted as a shot-noise, which is thus well-defined when $\int_{\mathbb{R}^2} |\psi_\lambda(x)| dx < \infty$. We denote the wavelet coefficients of ϕ by $\{\phi \star \psi_\lambda(x)\}_{\lambda \in \Lambda, x \in \mathbb{R}^2}$.

Regarding wavelet transforms of a counting measure $\mu \in \mathbb{M}^s$ observed in the finite window W_s, unless otherwise specified, we consider *periodic edge connection*; i.e., we use periodic wavelets defined at $x = (x_1, x_2) \in W_s$ with $\psi^s_\lambda(x_1, x_2) := \sum_{n_1 \in \mathbb{Z},} \sum_{n_2 \in \mathbb{Z}} \psi_\lambda(x_1 + 2sn_1, x_2 + 2sn_2)$. When there is no ambiguity, for $\mu \in \mathbb{M}^s$ we write

$$\mu \circledast \psi_\lambda(x) = \int_{W_s} \psi^s_\lambda(x - y) \mu(dy), \quad \lambda \in \Lambda. \tag{16}$$

We denote the wavelet coefficients of $\mu \in \mathbb{M}^s$ by $\{\mu \circledast \psi_\lambda(x)\}_{\lambda \in \Lambda, x \in W_s}$.

4.1.2 Wavelet phase harmonics

Phase harmonics [24] of a complex number $z \in \mathbb{C}$ are defined by multiplying its phase $\varphi(z)$ by integers k, while keeping the modulus constant, i.e.

$$\forall\, k \in \mathbb{Z},\ [z]^k := |z| e^{ik\varphi(z)}.$$

Note that $[z]^0 = |z|$, $[z]^1 = z$, and $[z]^{-1} = z^*$ (complex conjugate of z). More generally, $([z]^k)^* = [z]^{-k}$ and $|[z]^k| = |z|$ for $k \in \mathbb{Z}$.

We apply the phase harmonics to adjust the phase of the wavelet coefficients. For all $x \in \mathbb{R}^2, \lambda \in \Lambda$, $k \in \mathbb{Z}$, let's denote the *wavelet phase harmonics* of $\phi \in \mathbb{M}$ by

$$[\phi \star \psi_\lambda(x)]^k = |\phi \star \psi_\lambda(x)| e^{ik\varphi(\phi \star \psi_\lambda(x))}.$$

The phase of the wavelet coefficient $\varphi(\phi \star \psi_\lambda(x))$ is multiplied by k, whereas the modulus $|\phi \star \psi_\lambda(x)|$ remains the same for all k. The wavelet phase harmonics of $\bar{\phi} \in \mathbb{M}^s$ is defined similarly. Note that the wavelet phase harmonics at $k = 1$ are exactly the wavelet coefficients .

As illustrated in [36], when ϕ is a realization of a stationary process, the frequency support of $\phi \star \psi_\lambda$, which is centered around λ, is shifted and dilated by the phase harmonics. As a consequence, $[\phi \star \psi_\lambda]^k$ has a frequency support roughly centered around $k\lambda$. This non-linear frequency transposition property is crucial to capture dependencies of the wavelet coefficients across scales and angles. It is shown numerically in [24] that capturing such interactions allows one to recover signals with sparse wavelet coefficients. We observe similar reconstruction phenomenon in some point processes, as we shall illustrate in Section 6.

4.2 Wavelet phase harmonic covariance descriptors

As the wavelet transform is a linear transformation of Φ, it is known that the covariance between $\Phi \star \psi_\lambda(x)$ and $\Phi \star \psi_{\lambda'}(x')$ depends only on the mean intensity and second-order correlations of a stationary point process Φ, see e.g. [9, Eq. (5.27)], which only gives partial information on the process distribution. A classical way to capture dependencies between wavelet coefficients is to compute their higher order moments. However, as the order goes high, so does the variance of the moment estimator (which may violate (P1)). Based on the frequency transposition property of the phase harmonics (see Section 4.1.2), we shall explain how to capture dependencies between the wavelet coefficients at different locations and frequencies by computing the covariance between wavelet phase harmonics. Note that the wavelet phase harmonics does not increase the amplitude of the wavelet coefficients with $k > 1$. This approach may thus significantly reduce the variance of the descriptor K (to satisfy (P1)) compared to the higher order moments, while still capturing information beyond second-order correlations (to satisfy (P2)).

Let $\mathrm{Cov}(A, B) = \mathbb{E}[AB^*] - \mathbb{E}[A]E[B^*]$ denote the covariance between two complex random variables A and B. The wavelet phase harmonic covariance of Φ is defined by

$$\mathrm{Cov}([\Phi \star \psi_\lambda(x)]^k, [\Phi \star \psi_{\lambda'}(x')]^{k'}), \tag{17}$$

for paris of $(x, x') \in \mathbb{R}^2 \times \mathbb{R}^2$, $(\lambda, \lambda') \in \Lambda^2$, and $(k, k') \in \mathbb{Z}^2$.

In particular when $k \neq 1$ or $k' \neq 1$, the covariance measures the dependencies between the wavelet coefficients. As explained in [24], for a stationary process Φ, the overlap between the frequency support of $[\Phi \star \psi_\lambda]^k$ and that of $[\Phi \star \psi_{\lambda'}]^{k'}$ is necessary for the wavelet phase harmonic covariance to be large. Due to the frequency transposition property of the wavelet phase harmonics, it is empirically verified that the covariance at $k\lambda \approx k'\lambda'$ is often non negligible when the process is non-Gaussian (i.e. has structures beyond 2nd order correlations). We shall also follow this empirical rule to select a covariance set Γ_H to describe point processes.

Let $v_{\lambda,k} = \mathbb{E}([\Phi \star \psi_\lambda(x)]^k)$, we define the descriptors $K(\mu)$ using (10) with the functions $f_i(\mu)$ of $\mu \in W_s$ taken from a covariance set Γ_H (specified in detail in Section 4.2.1)

$$\left\{ \left([\mu \circledast \psi_\lambda(0)]^k - v_{\lambda,k} \right) \left([\mu \circledast \psi_{\lambda'}(-\tau')]^{k'} - v_{\lambda',k'} \right)^* \right\}_{i=(\lambda,k,\lambda',k',\tau') \in \Gamma_H}. \tag{18}$$

As Φ is stationary, (17) depends only on $x - x'$, it suffices to use the vectors τ' to measure the differences between x and x'.

Recall from (16), the wavelet transform of the point pattern $\mu \in W_s$ in (18) is taken with respect the the periodic variant of the wavelet ψ^s_λ, and taking the spatial average (10) gives the descriptor in the form [8]

$$K(\mu) = \left(\frac{1}{|W_s|} \int_{W_s} \left([\mu \circledast \psi_\lambda(x)]^k - v_{\lambda,k} \right) \left([\mu \circledast \psi_{\lambda'}(x - \tau')]^{k'} - v_{\lambda',k'} \right)^* dx \right)_{(\lambda,k,\lambda',k',\tau') \in \Gamma_H} \tag{19}$$

Note $K(\mu)$ is invariant with respect to any circular translation $\bar{S}_x$ of $\mu \in \mathbb{M}^s$ on $x \in W_s$.

In the numerical computation, we shall replace $v_{\lambda,k}$ in (18) by $\bar{v}_{\lambda,k} = \frac{1}{|W_s|} \int_{W_s} [\bar{\phi} \circledast \psi_\lambda(x)]^k dx$ as a plug-in estimator for the first-order moment $v_{\lambda,k}$. The $K(\bar{\phi})$ modified in this way becomes an empirical estimator of the covariances in (17). This is a good approximation of $K(\mu)$ as the estimation variance of the covariance moments is typically much larger than the first-order moments.

4.2.1 Choice of the covariance set Γ_H

To capture dependencies of the wavelet coefficients across scales, angles and spatial locations, we are going to select non-negligible wavelet phase harmonic covariances, defined a set Γ_H formed by the pairs of (λ, λ') and (k, k'), as well as τ'. Due to the symmetries in the wavelet coefficients [36], it is sufficient to consider $k \geq 0$ and $k' \geq 0$. Different from [36], we propose a more intuitive way to choose τ' so as to capture local structures of edges that we observe in point processes.

Choice of τ' When the wavelet coefficient $\Phi \star \psi_\lambda(x)$ has a large amplitude, it detects edges in Φ oriented in the direction of ψ_λ near x (in which ψ_λ oscillates the most). In the orthogonal direction e_θ, the wavelet ψ_λ has a smooth bump-like shape. More precisely, if ψ_λ has central frequency $r_\theta \xi_0$, e_θ is a vector defined by $e_\theta := \frac{C}{|\xi_0|} r_{\theta + \frac{\pi}{2}} \xi_0$ (recall that C is such that $|\psi(x)|$ is negligible if $|x| > C$). We shall choose either $\tau' = 0$ or $\tau' = 2^{j'} e_\theta$. For the non-zero τ', it corresponds to a shift of the wavelet in its smooth direction e_θ. This allows us to capture correlations between nearby edges, one located at $x' = x - \tau'$ and the other at x. When θ and θ' are different, the covariance can be interpreted as a measure of local curvatures along nearby edges. Empirically, we observe that including the non-zero τ' in the descriptor improves the geometrical structures in the samples of the model (based on our visual evaluation in Section 6).

Full list of Γ_H We specify the Γ_H whose total number of elements is at most $O(L^2 J^2)$. For any $\tau' \in \{0, 2^{j'} e_\theta\}$, we include the following set in Γ_H for $\lambda = 2^{-j} r_{-\theta} \xi_0, \lambda' = 2^{-j'} r_{-\theta'} \xi_0$ for $0 \leq j \leq j' < J$.

Due to the Hermitian symmetry in any covariance matrix, we also limit $0 \leq k \leq k'$. In addition, we include $\lambda = \lambda' = 0$ (exclusively for $j = j', k = k' = 1, \tau' = 0$) to capture the 2nd-order correlations at scales equal or larger than J.

For $j = j'$, we want to capture 2nd-order correlations, as well as wavelet dependencies at different orientations:

- $k = k' = 0$ or $k = 0, k' = 1$, for all angle pairs $(\theta, \theta') \in \frac{2\pi}{L} [|0, L-1|]^2$. They capture dependencies between different orientations, without and with phase information.
- $k = k' = 1$, with angle pairs restricted to $|\theta' - \theta| \leq 4\pi/L$. They capture 2nd-order correlations of the process Φ.

For $j \neq j'$, we aim at capturing wavelet dependencies between different scales and orientations. We limit the scales difference to $j' \leq j + \Delta_j$, and choose $\Delta_j = 2$ [9]:

[8] Ideally, we also aim to match the 1st order moments $v_{\lambda,k}$. However, it is a hard optimization problem when using the quadratic energy $E_{\bar{\phi}}(\mu)$ in (9), to simultaneously match the 1st and 2nd order moments. This is why we include them indirectly in the K in (19), as done in [24, 36].

[9] Our choice for the value of Δ_j is a trade-off between the variance and the discriminate power of K. For a study of the impact of Δ_j on the latter, we refer the reader to [1].

- $k = 0$, $k' = 0, 1, 2$, for all angle pairs $(\theta, \theta') \in [0, 2\pi)^2$. They capture the dependencies at different scales and orientations, without and with phase information.
- $k = 1$, $k' = 2^{j'-j}$, angles $|\theta' - \theta| \leq 4\pi/L$ for the dependencies at different scales, plus support superimposition between orientations.

4.2.2 Variants of particle descent models

The choice of K has a direct impact on the model ability to approximate the distribution of Φ. Two situations must be avoided: 1) K carries too little information about Φ ((P2) does not hold), or 2) K carries wrong information ((P1) does not hold). In this section, we discuss these situations in the context of the wavelet phase harmonic covariance descriptor (defined in (19)).

The covariance set Γ_H (and hence the descriptor K) depends on the parameter J, which is the maximal scale of the wavelet transform. Intuitively, the largest structure of the point pattern we can capture through the descriptor $K(\mu)$ in (19) has diameter $2^J C$. Recall that C is the diameter of the "effective" support of ψ (so that $\psi(x) \simeq 0$ if $|x| > C$). In other words, $K(\mu)$ can capture interactions of the points of μ within a distance smaller than $2^J C$.

Ergodic model To avoid situation 1), we must choose the maximal scale parameter J large enough to capture sufficient structural information about Φ. However, J must not be too large, in order for all the moments to be well estimated ((P1) must be satisfied, otherwise we fall in situation 2)). A suitable choice for the maximal scale parameter J would be one allowing for a good trade-off between satisfying the sufficiency of K, while maintaining the concentration property. We call our descriptor with such a J the *ergodic descriptor*, and the corresponding model the *ergodic model*, as they rely on the ergodic property of the underlying process.

From reconstruction to the regularized model The ergodic descriptor may be too limited if Φ exhibits structures at scales much larger than the largest parameter J such that K satisfies (P1). The ergodic model could thus fail to satisfy (P2), and the model would fail to approximate $\mathcal{L}_\Phi$ in a satisfying way (this would depend on the evaluation method, refer to Section 6 for the evaluation of our models, and Figure 4 for an example of this situation). This motives us to increase the size of the descriptors, by increasing J up to $O(\log_2 s)$. However, as J increases, so does the variance of K, and therefore (P1) would be violated. This problem is well illustrated in the case of our descriptor (19), where the model can memorize $\bar{\phi}$ (i.e. sampling from the model amounts to generate translated versions of $\bar{\phi}$).[10] Hence, we are facing a problem where increasing the descriptor size is necessary to have a good distribution approximation, but it will violate (P1).

To address this issue, we propose an alternative model allowing for the use of a J that may not satisfy (P1). It consists in adding a regularization term to the energy $E_{\bar{\phi}}(\mu)$ in (9), so as to prevent the gradient descent algorithm converge to the global minimum which is $\bar{\phi}$. This can be realized, for example, using a Wassertein-type distance between the current and the initial configuration of points, as described in Appendix A. Note that in this *regularized model* the invariance properties of Theorem 3.3 still hold.

Non-periodic integration (not evaluated in this paper) Recall that the descriptor $K(\mu)$ defined in (19) applies periodic boundary correction to the data $\mu \in W_s$ by using periodized wavelets. Thus, it is invariant with respect to any translation $\bar{S}_x$ of $\mu \in \mathbb{M}^s$ on $x \in W_s$. By Theorem 3.3, our model (as described in Section 3 or with regularisation presented in Appendix A) with Poisson initial condition $\bar{\Phi}_0$ (conditioned to have $\bar{\phi}(W_s)$ points) generates circular-stationary point processes $\bar{\Phi}_n$ on W_s (i.e. having distribution invariant with respect to translation $\bar{S}_x$ in W_s). As already pointed out in Section 3.2, this may be seen as a desirable property of the generative model (it increases the diversity and may be seen as an ersatz of the stationarity of the original, approximated distribution). If it is not the case, or if the bias[11] introduced by the

[10] Note that the violation of (P1) does not imply memorizing $\bar{\phi}$. While this is the case for our descriptor when increasing J, this is for instance not the case when using Fourier spectrum descriptors, as illustrated in C.1

[11] The bias of the estimator $\bar{v}_{\lambda,k}$ is non-zero if $\mathbb{E}([\Phi \star \psi_\lambda(x)]^k) \neq \mathbb{E}([\bar{\Phi} \circledast \psi_\lambda(x)]^k)$ for x being close to the border of the window W_s.

periodic boundary correction is too big, one may want to deal with the boundary problem in a different way, for example considering a non-periodic integrals in (19) over some smaller window. In particular, we suggest a *scale-dependent reduction of the integration window*, pertinent when the wavelet ψ has a compact (or approximately compact) spatial support. Specifically, we consider a new descriptor $\tilde{K}$ by considering the integrals in (19) with $i = (\lambda, k, \lambda', k', \tau') \in \Gamma_H$ over smaller windows $W_{s_i} \subset W_s$, such that boundary effects are neglibible.

Note that, while the above scale-dependent reduction of the averaging window removes the bias related to the periodic integration over the entire observation window W_s, it potentially increases the variance of $\tilde{K}(\bar{\Phi})$. Note also, the descriptor $\tilde{K}$ is no longer translation invariant, so the model derived by it is not circular stationary. We call this non-periodic integration. It will be not evaluated in this paper.

5 Numerical scheme with multiscale optimization

In this section we discuss some details of the implementation of the algorithm presented in Section 3 with the descriptor based on the wavelet phase harmonics presented in Section 4. These details are, in order:

- *Discretization* for an approximate calculation of the covariance of the wavelet phase harmonics: necessary to accelerate the calculation of the descriptor and the gradients.
- *Multiscale optimization*: allowing one to avoid shallow local minima in the gradient descent model.
- Final randomization (*blurring*): to get rid of some clusterisation (clumping) artifact caused by the initial discretization.

5.1 Discretization

5.1.1 Differentiable discretization of atomic measures

To compute the descriptor K in (19) for a point measure μ, we need to integrate functions over the observation window W_s (first for the convolution operators, then for the averages). Computationally efficient integration requires discretization of the atomic measure. The main difficulty is to do it in such a way that the (periodic) convolutions of the discretized atomic measures with wavelets, as in (16), remain differentiable with respect to the positions of the original atoms in μ, so that we can still perform gradient descent. Classical finite element methods may not achieve this goal efficiently.

We are going to approximate our atomic measures on W_s by matrices (images) of given size $N \times N$ (the image resolution), and then use an automatic differentiation software [2] to perform the following operations.[12] We first map a given point measure μ on W_s to a continuous function by the convolution

$$\mu_\sigma := \mu \circledast g_\sigma \tag{20}$$

with a (periodized) Gaussian function g_σ of given standard deviation σ. Then we evaluate μ_σ on a $N \times N$ regular grid inside W_s and denote the resulting matrix μ_σ^N. The convolution with a Gaussian function makes each entry of μ_σ^N smoothly depends on the atom positions of μ. We then compute $\bar{K}(\mu_\sigma^N)$ instead of $K(\mu)$, where $\bar{K}$ is this discrete analogy of the descriptor (19)[13] (cf. [36]). The gradient of the energy $|\bar{K}(\mu_\sigma^N) - \bar{K}(\bar{\phi}_\sigma^N)|$ with respect to each atom position of μ can therefore be computed using the automatic differentiation.

The Gaussian function are low-pass frequency filters. It is needed to cut-off high frequency information of μ so that μ_σ can be discretized into an image with negligible alisaing effect. This means that μ_σ carries the information on the positions of μ up to some precision which depends on σ. The subsequent evaluation of μ_σ on the grid $N \times N$ in μ_σ^N implies that σ cannot be taken too small. Indeed, we take $\sigma_{\min} = \frac{s}{N}$ as the lowest value of σ.

[12] The code to reproduce the results is run on GPU to take advantage of parallel computations.

[13] This discretization makes our descriptor only invariant to discrete translations, as the value of a pixel continuously depends on the positions of the atoms.

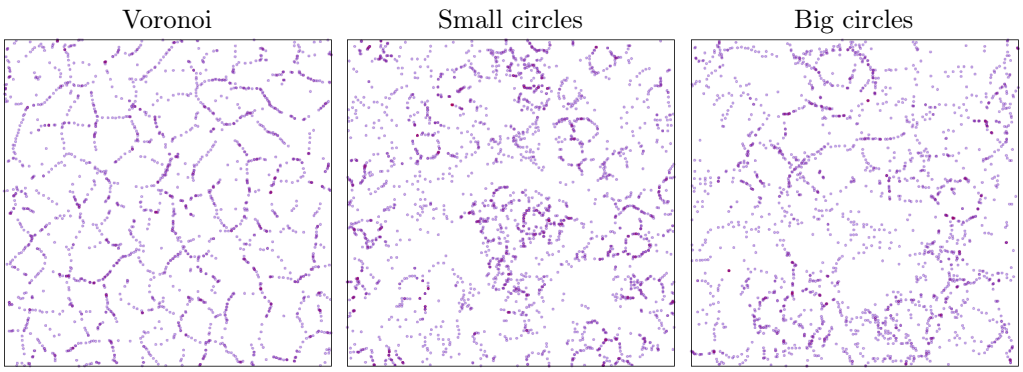

Figure 2: An attempt of the reconstruction (i.e., a realization generated with the full-scale descriptor) without multi-scale optimization of the Voronoi, Small circle, and Big circle original realization presented on Figure 4. For the Big Circle case, the loss (the value $E_{\bar{\phi}}(\bar{\phi}_n)$) is equal to 3.6269, while the loss of the reconstruction with multi-scale optimization (also presented on Figure 4) goes down to 0.0241 (a reduction of 99 percent).

5.1.2 Wavelet discretization

As stated in section 4.1.1, the family of wavelets used in our descriptor is constructed by dilating the mother wavelet ψ in the range of the scales $0 \leq j < J$. Based on the choice of $\sigma_{\min}$, we set $C = \frac{2s}{N}$. In this way, the spatial support of ψ has a radius C of one pixel of the image. The choice of J can be decided based on the visual structures in the observation. For example, if we want to model structures whose spatial size is close to the size of the window $[0, 1/8]^2 \subset W_s$, we shall set $2^J C = 1/8$, i.e. $J = \log_2(N) - 3 - \log_2(2s)$.

5.2 Multiscale optimization

Phase harmonic covariance moments of point process images (i.e. point patterns converted into regular pixel grids, as described above) may have large values at high frequencies (values of λ, λ' in (19) with small j, j'), due to the fact that the point-images are very sparse and highly discontinuous. This implies that these high frequency statistics have an important impact on the gradient of K, which in turn can lead to the gradient descent model being trapped at shallow local minima, where only the high frequencies are well optimized to match the observation.

This phenomenon arises in particular when our descriptor K does not satisfy the concentration property, and characterize the sample $\bar{\phi}$ rather than the underlying distribution (cf. Section 4.2.2). In this case, the particle gradient descent algorithm often fails to reach Ω_ϵ, and we observe that the generated samples fail to match the low frequency elements of K. Figure 2 shows an attempt to reconstruct samples from 3 different Cox processes (see Section 6.1.1) without multiscale optimization, and a reconstruction with it.

This can be overcome by matching the descriptors from low frequency to high frequency in a sequential order, through an appropriate modulation of the parameter $\sigma \in \Sigma = (\sigma_0, \sigma_1, ..., \sigma_{J-1})$ of the Gaussian functions used to discretize μ, introduced in Section 5.1.

Indeed, since Gaussian functions are low-pass filters, we can interpret the convolution in (20) as a blurring, limiting the space localization of Dirac measures. When such smoothing of the point pattern is done by a Gaussian function that has a large variance, the high frequencies of the signal function are close to 0 and the same holds true for the phase harmonics, because wavelets are localized in frequency. Therefore the wavelet phase harmonics are dominated by the low frequencies. Thus, by smoothing the observed sample and generating the optimal one with high variance Gaussian function, we create a new objective leading, in the gradient descent optimization, to a point configuration for which only low frequencies moments are matched with the ones of our observed sample. We propose thus a *multiscale* gradient descent procedure that

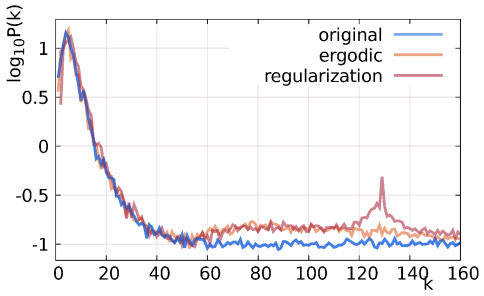

Figure 3: Power spectrum plots showing clustering at higher frequencies (k larger than 60) — artefact of the discretization of the point configurations in the generative model (appearing both in the ergodic and the regularized version). This effect can be removed by appropriate final blurring, cf. Figure 7.

consists in choosing first a high value for precision parameter σ, run the optimization algorithm, and then reduce the value of σ to run the optimization again, starting from the result of the previous run (and repeat this operation several times).

In more details, let N be the resolution of our discretization (in what follows, we shall drop the dependency in N for notations simplicity). We choose $\sigma_j := 2^{J-j-2}$. Starting with σ_0, we calculate the descriptor $\bar{K}(\bar{\phi}_{\sigma_0})$ of the given point pattern $\bar{\phi}$ and of an initial Poisson configuration $\bar{\phi}_0$ on W_s, conditioned on having $\bar{\phi}(W_s)$ points. We denote the corresponding energy $E^{\sigma_0}_{\bar{\phi}}$. For numerical efficiency, we perform the gradient descent procedure using the L-BFGS optimization algorithm [20] to optimize the corresponding energy function $E^{\sigma_0}_{\bar{\phi}}$. The obtained optimal solution are denoted by $\bar{\phi}_1$. We then repeat this operation recursively starting from $\bar{\phi}_j$ using precision σ_j for $1 < j < J$. The whole, multiscale optimization can be summarized as follows:

- Sample $\bar{\phi}_0$ from Poisson distribution on W_s given $\bar{\phi}(W_s)$ points.
- For $j = 0 \ldots, J-1$:
  - Run optimization on $\bar{\phi}_{j+1}$ to minimize $E^{\sigma_j}_{\bar{\phi}} := \frac{1}{2}||\bar{K}((\bar{\phi}_{j+1})_{\sigma_j}) - \bar{K}(\bar{\phi}_{\sigma_j})||^2$
- Return $\bar{\phi}_J$.

Empirical evidence shows that the above multi-scale optimization procedure allows one to reconstruct (modulo translation) the observed sample when using full-scale wavelet phase harmonics, which is not the case when simultaneously optimizing all frequencies. In order to preserve the ability to reproduce geometric structures at all scales, we shall apply this multiscale optimization method to all the variants of our model.

5.3 Final blurring

Note that our discretization μ^N_σ of μ does *not* make us entirely lose the information at scales smaller than 1/N (which would be the case if we were approximating μ by $(N \times N)$-pixel image by counting the number of points within each the pixel square of size $2s/N \times 2s/N$). Indeed, the entries (pixels) of $\mu_{\sigma,N}$ continuously depend on the exact positions of the points. We are then able to (at least partially) retrieve information lost by the mapping from positions to images. This information is of course most significant for the the smallest precision parameter $\sigma = \sigma_{\min}$ applied in the last step of our multiscale optimization.

However, we observe that the contrast between the continuous nature of our objects and the discrete approximation of the moments we use creates undesired artificial clustering structures at frequencies higher than the image resolution. Furthermore, the model with the regularization (mentioned at the end of Section 4.2.2 and described in Appendix A) creates yet another clustering effect where the points tend to cluster in a grid-like structure. We explain this latter phenomenon by the fact that, while moments computed in a discrete fashion are invariant to small displacement of points, the regularization function is not. Therefore, when put in an optimal position w.r.t. the descriptor K, the points are drawn to their corresponding points in the initial configuration, and

move as close as possible to them (without modifying too much K), that is to the edges of the pixel squares. We show both clustering effects in Figure 3. For details regarding the power spectrum, see Section 6.2.

To remove this artificial clustering, which we believe is an artifact of the discretization causing a loss of high-frequency information, we chose to force these high frequencies to be "as random as possible", i.e. to have Poisson-like structure. To this end, we introduce a uniform i.i.d. perturbation of the positions of points after the last optimization run. It can be viewed as an additional, this time stochastic, measure transport, following the deterministic one from the particle gradient descent.

This final randomization can be viewed as enforcing a-priori information on high frequency structures of the process: Poisson-like structure. If we know that this is not the case (e.g. hardcore repulsive processes), such randomization should not be applied.

6 Numerical experiments

In this section we present some numeral experiments involving our generative model. We begin by presenting in Section 6.1 our numerical settings, in particular the distributions of point processes whose samples are used as original point patterns. We next evaluate how well our generative model with the phase harmonic covariance descriptor can generate samples similar to those given by the original point processes. In Section 6.2 we first present the *reconstruction, ergodic and regularized* models of the original processes, discussed in Section 4.2.2. We first evaluate these models by comparing samples to original samples, visually as well as by calculating their power spectra.

In order to further evaluate qualitatively how well our model captures visual geometric structures, and to gain some insight into the ability of our models to produce diverse samples, we shall use the topological data analysis (TDA), derived from the theory of persistent homology. The comparison will be done in Section 6.3. The results regarding the regularized version of our model (cf. Section 4.2.2 and details in Appendix A) are presented as well.

6.1 Numerical settings

We first describe the original point processes that we shall evaluate the particle gradient-descent model, then specify the parameters of the model in the numerical experiments.

6.1.1 Original point process models

For our experiments, we choose point process distributions that show complex geometric structures, for which we can easily recognize visual features that are not fully captured by second order correlations (2nd order correlations basically capture pairwise "interactions" between points e.g. clustering or repulsion).

We begin by presenting results for *Cox (double-stochastic Poisson) processes* with Poisson points living on one dimensional structures generated by two famous stochastic geometric models, namely edges of the Voronoi tessellation, and the Boolean model with circular grains of fixed radius considered in [12, Example 10.6]. Both underlying geometric models are generated by a Poisson parent process within the observation window W_s, and we construct these models in a periodic way to avoid border effects. We call the respective Cox processes *Voronoi* and *Circle* processes. Note that, for these two processes, Poisson points live on different geometric shapes: polygons for the Voronoi and possibly overlapping circles for other one. Additionally, we consider two different radii of circles.

Secondly, we take interest in distributions having *turbulent intensity* (derived from the simulations of a periodic turbulent flow driven by 2d Navier-Stokes equations [29]). These intensities exhibit complex mulsticale structures, and are representations of physical phenomena, known to be difficult to model faithfully. Furthermore, the distributions we consider have much greater intensities that the previous Cox models. We consider two different turbulent intensities: one with limited range correlations (cf. Figure 5), and one with larger scale structures (cf. Figure 6). From the former, we sample a Poisson point process. From the latter, we sample three different processes, exhibiting distinct microscopic structures (repulsive, independent or clustering): a Matern

Cox			Turbulent intensity				
Voronoi	Small circles	Big circles	Poisson I	Poisson II	Poisson III	Hard core	Cluster
1 900	2 500	2 000	40 000	19 000	3 800	1 700	13 000

Table 1: Approximate number of points in realizations from different models.

cluster process, a Poisson point process and a Matern II *hard-core* process, see [12, Example 5.5 and Section 5.4, respectively]. We study the ability of our model to reproduce simultaneously the macroscopic (i.e. the turbulent intensity) and microscopic structures (i.e. at small scales) of the process.

All processes are sampled on $W_s = (-1/2, 1/2]$, i.e. $s = \frac{1}{2}$. Table 1 presents approximate numbers of points in the respective models. Note that, for comparison, point patterns considered in [33] have around 400 points.

6.1.2 Image resolution, number of iterations and computation time.

As discussed in Section 5.1, point configurations are convoluted with Gaussian densities and evaluated on $N \times N$ grid (as images) in order to efficiently compute our descriptors, and move the particles with gradient descent. The ultimate Gaussian variance (precision) of this mapping is $\sigma_{\min} = \frac{1}{2N}$. The larger N is, the more information we are able to keep (in high frequencies), but the larger the computation time. We chose for our experiments a resolution of $N = 128$. We show an example where a higher resolution, $N = 256$, is used to capture most of the high frequency information. The number of iterations of the algorithm is chosen to be 100 for each σ_j (a total of 400 iterations for $N = 128$, and 500 iterations for $N = 256$).

For the ergodic model, we take $J = \log(N) - 3$. For the reconstruction and regularized models, we take the set Γ_H with the largest scale parameter $J = \log_2(N) - 2$. In both cases, the number of angles in the steerable wavelets is $L = 8$.

The average computation time on 4 modern GPU[14] for a synthesis of the ergodic model for the turbulent process (having roughly 19 000 points) with resolution $N = 256$ is between 5 and 10 minutes while the same task at the resolution $N = 128$ takes between 1 and 2 minutes.

6.2 Visual evaluation and spectrum comparison

We evaluate the ability of our model to capture and reproduce geometric structures exhibited by realizations of the point processes described in Section 6.1. A natural first method to assess the sufficiency of a generative model is visual evaluation, which is widely used in image analysis but subjective. To make the evaluation more objective, we evaluate second order correlations. We estimate the power spectra of the original processes and that of our models, which are statistics approximating the Bartlett spectrum of the point processes distributions (cf. Appendix C).

To perform the statistical evaluation, we generate (for each original distribution) 10 i.i.d. syntheses from the same model at resolution $N = 128$ (i.e. from the same observation sample $\bar{\phi}$, but with different initial configurations $\bar{\phi}_0$). We average the power spectra of the 10 syntheses, and compare it to the average of 10 i.i.d. samples from the true distribution. All these samples will also serve in Section 6.3 to compare their geometric similarities.

Figure 4 shows a study of three Cox models. The first line presents samples from the original distributions. The second line presents samples from the model using the full scale descriptor. As discussed in Section 4.2.2, K does not concentrate in this setup, and the result is the memorization of the observation sample $\bar{\phi}$. Indeed, the second line of Figure 4 shows quite faithful approximations of the original samples subjected to a periodic translation. Observe the reconstruction is accurate (up to some precision error) including the details (points on the geometric structures).

The third line of Figure 4 shows realizations sampled using the ergodic model for different original distributions. As discussed in Section 3, in order to get some diversity in our model, we need to choose a descriptor K that respects (P1). This could be achieved by choosing the maximal

[14]Hardware: Tesla P100-PCIE-16GB GPU

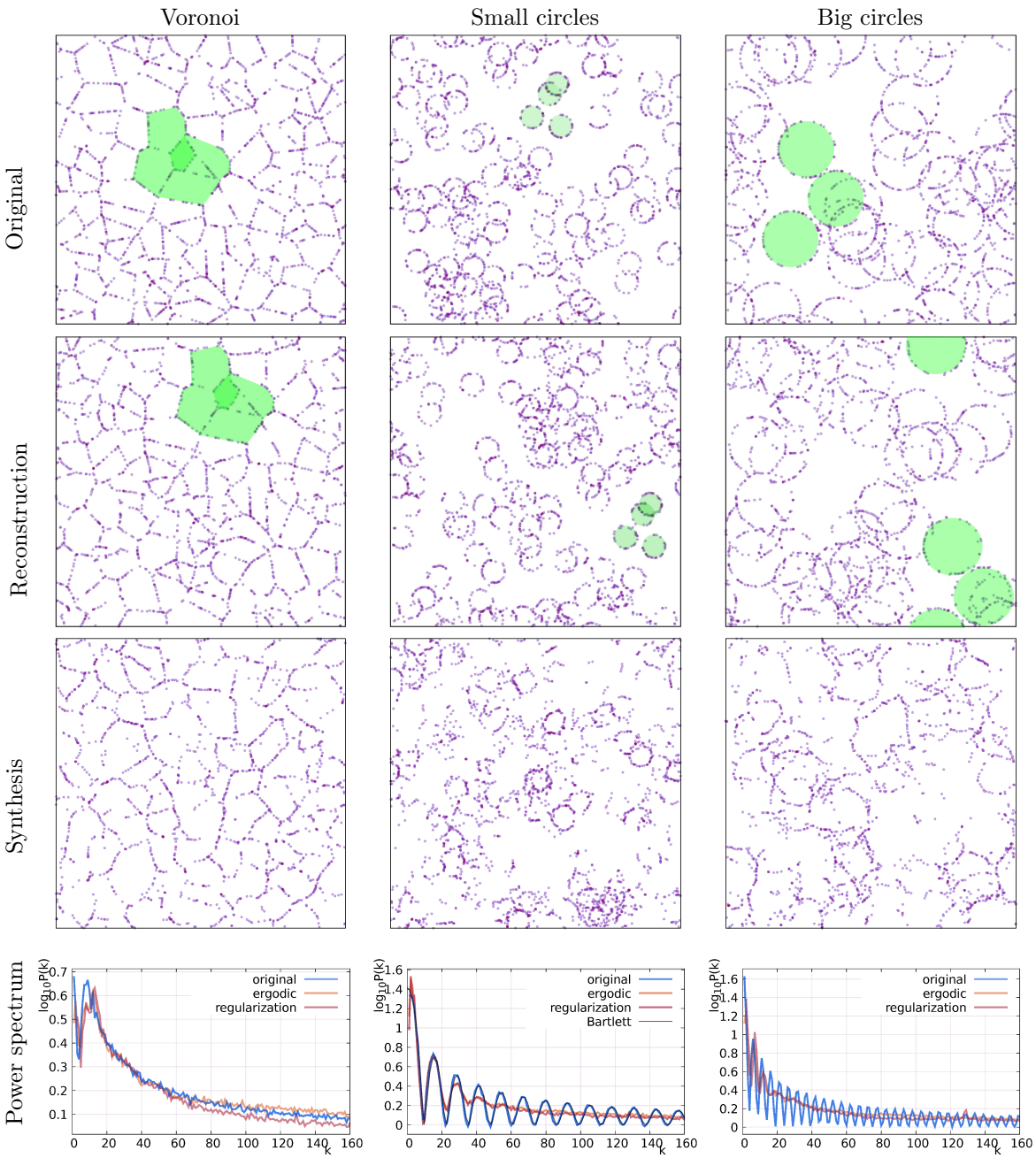

Figure 4: Three Cox models; original sample, reconstruction with the full scale descriptor and synthesis with ergodic descriptor. Power spectrum plots averaged over 10 realisations.

scale J that is adapted to the largest size of the structures that we want to model. Based on the visual observations and our analysis in Section 5.1.2, we have set $J = \log_2(N) - 3$ to model structures whose spatial size is at most $1/8$ of the window W_s, for $s = 1/2$. Observe, the geometric shapes (polygons, circles) of larger scales are not so well reproduced as in the reconstruction, more particularly so in the Big Circles case. This is due to the missing large scale (non-ergodic) descriptor components.

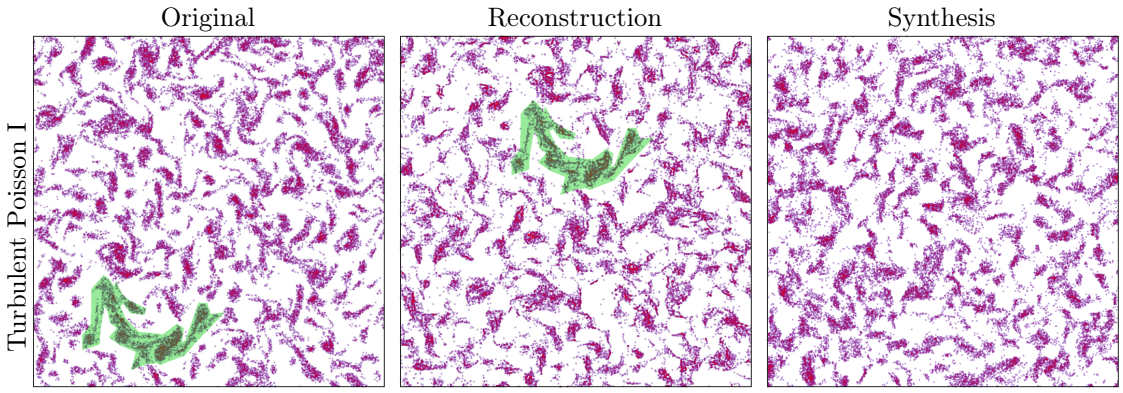

Figure 5: Large scale turbulent Poisson I point process (about 40 000 points): original sample, reconstruction with the full scale descriptor and synthesis with ergodic descriptor.

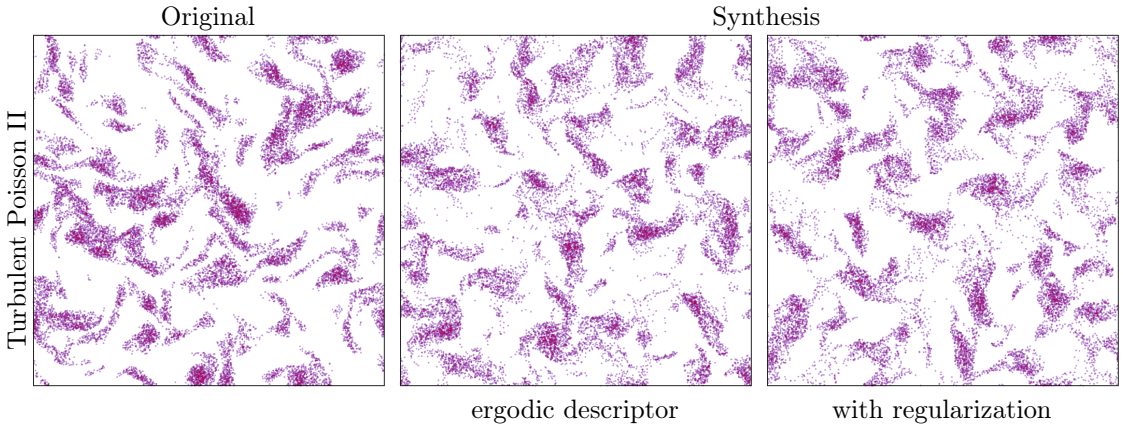

Figure 6: Turbulent Poisson II point process (about 19 000 points): original sample and two samples generated using the ergodic descriptor as well as the full scale descriptor with regularization; cf. Appendix A.

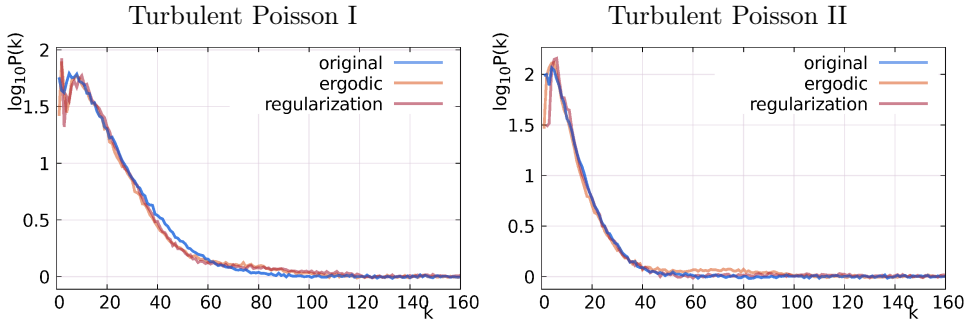

Figure 7: Power spectrum plots averaged over 10 realisations for Turbulent Poisson I and II.

In the last line of Figure 4, we present the power spectrum $\tilde{U}_k(\mu)$, for $k \in \mathbb{N} \cap [1, 128[$ (cf. Appendix C) from the original distributions and as well as our two models (cf. Appendix A). Point configurations generated from the regularized model cannot be visually discriminated from the samples of the ergodic model (as illustrated below in Figure 6 for the Poisson II turbulent process). Some small differences can be however observed in the spectrum, and via the TDA,

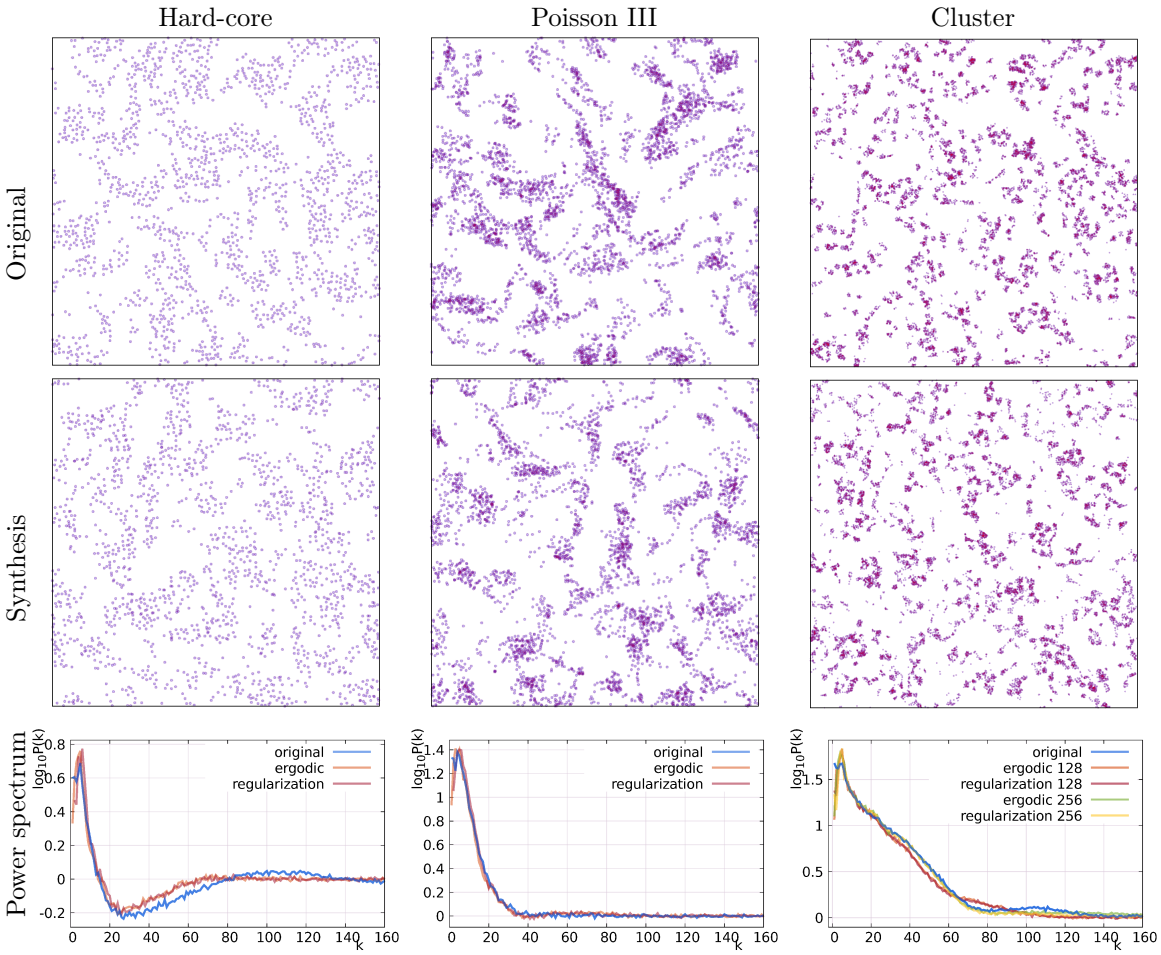

Figure 8: Synthesis of turbulence images with various microscopic structures and their power spectrum plots . The spectrum of the cluster model (lower right figure) was analysed when generated at the (standard for this paper) resolution $N = 128$ and a higher one $N = 256$, allowing for higher frequencies reproduction (the clusters are more concentrated, as in the original image). The presented synthesis of the Cluster model is at resolution $N = 256$.

see Section 6.3.) The norm of the largest frequencies captured by our discretized descriptor $\bar{K}$ (cf. Section 5.1) at the resolution $N = 128$ equals approximately to $k = 50$. Observe that for point processes having sufficient smooth spectrum (e.g. the Voronoi Cox model on Figure 4 and Turbulent Poisson II model presented on Figures 6) our generative model reproduces relatively well the spectrum up to this frequency threshold. However, because the wavelet convolutions average the spectral information over frequency bands [36], our descriptor does not capture fast oscillations in the power spectrum. This phenomenon is well illustrated in the cases of Small and Big Circles Cox processes. For the Small Circle model (last line, middle plot of Figure 4) we show also the theoretical value of the Bartlett spectrum calculated via expression (33) in appendix. It well matches the power spectrum estimated form the original model realizations thus confirming (32) in Appendix C.2.

Figure 5 presents a similar analysis (original sample, reconstruction with the full-scale descriptor, ergodic synthesis) for Turbulent Poisson model I. Note that this reconstruction is not due to an overparametrization of the problem (i.e. we do not impose more constraints than degrees of freedom) as K has size about 15 000, compared to roughly 80 000 degrees of freedom (the two

coordinates of each point). Figure 6 presents an original sample, a synthesis from the ergodic model, and a synthesis from the regularized model for Turbulent Poisson model II (with about 19 000 points). We observe that the two models are visually similar for this distribution. Figure 7 presents the spectrum analysis for the two Turbulent Poisson models presented on Figures 5 and 6. The smooth elevation of the spectrum of Poisson II model (left plot) around $k = 70$ in the ergodic model is (we believe) a discretization artifact (cf. Section 5.3 for more details on this effect).

All models discussed up to now have non-correlated points at the microscopic level. Indeed, they are Cox processes, with Poisson (hence independent) points sitting on some random macroscopic structures. Figure 8 presents our analysis of three turbulent point process models having different microscopic structure: a hard-core, non-correlated (Poissonian) and a clustering one. We see that our generated samples (both with the ergodic descriptor and the regularized, full-scale one) capture to some extent this microscopic structure. For the clustering model, the presented synthesis is done with $N = 256$. In this case, we present also the spectrum analysis of samples generated using the resolution $N = 128$. For $N = 256$, the power spectrum is well reproduced up to the frequency about or $k = 80$?.

6.3 Persistent homology and topology analysis

As previously mentioned, power spectrum evaluation corresponds to the comparison of second order moments, which only partially capture geometric structures. Visual evaluation can be more discriminate (cf. Appendix C.1, Figure 11), but is subjective. To evaluate more precisely the ability of our model to capture the geometric structures of the given distributions, we shall use a representation of objects derived from persistent homology theory, which is a powerful algebraic tool for studying the topological structure of shapes, functions, or in our case point clouds. We shall perform this evaluation by comparing the *persistence diagrams* of the generated samples to those of the original ones. Furthermore, this representation allows us to evaluate (to some extent) how distinct samples from our model are from one another.

We begin by a brief, intuitive presentation of persistence diagrams, and the whole comparison method that will be simply referred to as topology data analysis (TDA). For more details we refer the reader to [6, Section 11.5]. We then present the TDA of our point process distributions and models. TDA can be seen as a complementary tool with respect to the spectrum analysis, being more consistent with visual perception (see discussion in Appendix C.3).

Persistence diagram Persistent homology theory describes a way to encode the topological structure of a point cloud through a representation called *persistent diagram* (PD). It is constructed, for a given point configuration $\phi \in \mathbb{M}^s$, from the family $(G_r)_{r\geq 0}$ of Gilbert graphs, where the vertices are the positions of atoms of ϕ, and the edges are pairs of points closer to each other than r. (In our case we use the periodic metric.) Then, we fill-in the triangles (triplets of points joined by edges) of the graph. Points, edges and filled-in triangles constitute the so-called 2-skeleton of the Vietoris-Rips (VR) complex. For any $r \geq 0$, we study two characteristics of the skeleton: its *connected components*, and its *holes* (this latter notion is well formalized in the algebraic topology, in our case they correspond to the natural idea of a hole). Each connected component "is born" at time (radius) $r = 0$ and it "dies" at some time $r > 0$ when it is merged with another connected component. Similarly, each hole has a birth time ($r > 0$) corresponding to the minimal radius at which it appears, and a (larger) death time corresponding to the minimal radius for which the hole is completely filled-in by the triangles. The persistence diagram of ϕ is the collection of pairs of birth and death times of the connected components and holes. It is hence a point process in the positive orthant of the plane offering a multiscale (as our wavelet-base descriptor!) description of the topology of ϕ. As our descriptor, it is also stable to the small deformations of ϕ. It is hence interesting to use this alternative tool to evaluate our generative model.

Topological data analysis Our approach in this matter is inspired by [11], and we refer the reader to this paper for a more detailed description.

In order to compare the distribution of our models to the original distributions, we compute the PDs of our samples from each distribution (cf. Section 6.2 for a description of these samples).

Recall, these PDs can be viewed again as point clouds in two dimensions. Therefore, a distance between two PDs can be computed, and we use in this regard a periodic version of the Wasserstein distance [15] between two point clouds on the plane. We obtain in this way a distance matrix between different PDs (reflecting topological similarities or differences of the point processees realizations for which PDs were calculated.) We then apply a standard dimension reduction algorithm (namely Multi Dimensional Scaling) to this distance matrix, to represent every PD (and hence the corresponding sample) as one point on the plane, and we and visualize the representation of all samples.

TDA of our experiments In all plots in the first two lines of Figure 9 we study separately the Cox process and the turbulent models with different microscopic structures. In each plot we observe 30 *dots* (having different shapes) each representing one entire configuration of points in W_s. (The term "dot" is used to avoid confusion with points in W_s.) For each model there are 10 dots representing i.i.d. realizations of the original distribution, 10 representing realizations from the generative model with the ergodic descriptor, and 10 from this model with the regularized full-scale descriptor. We see in the case of the turbulent models that all 30 dots are relatively well mixed up, which can be interpreted as indicating good distributional approximation of both our generative models in this case. For the Cox point processes, it is easy to discriminate between the original and approximated point patterns. This is not a surprise, as the visual inspection has allowed us to state it more directly. The approximations of the Voronoi model exhibits also a smaller diversity (dots are more concentrated than the original realizations). For the Big Circles model, the regularized model performs better than the ergodic one (the dots of the regularized model are closer to the ones of the true distribution). The last line presents the TDA jointly for all Cox and Turbulent models, showing a good separation between the point processes we have considered.

We owe the reader the following additional explanation regarding our TDA: We were using R packages `TDAstats` [34] to calculate the PDs of our point patterns and `TDA` [17] to calculate their Wasserstein distances. None of these two packages allowed us (due to memory constraints) to treat realisations having significantly more than 2000 points. For this reason our analysis TDA of the Small circles, Poisson III and Cluster model is in fact done using random thinning of the given realisations (more precisely, random choice of a subset of 2000 points). This operation introduces possible artificial improvement of distribution approximation, as random thinning with small retention probability (and intensity preserving rescaling) makes any process converge in distribution to the Poisson process. In order to analyse the impact of thinning on our TDA, on Figure 10 we compare the original distribution and our ergodic model with 10 realisations of the same (original) point pattern subjected only to independent thinning. Our conclusion is that this thinning procedure does not significantly perturb the analysis of the concerned models.

Acknowledgements: This work was partly supported by the PRAIRIE 3IA Institute of the French ANR-19-P3IA-0001 program. Sixin Zhang is currently supported by the European Research Council (ERC FACTORY-CoG-6681839). Part of this work was done when Sixin Zhang was a postdoctoral researcher at ENS Paris, France.

References

[1] E. Allys, T. Marchand, J.-F. Cardoso, F. Villaescusa-Navarro, S. Ho, and S. Mallat. New interpretable statistics for large scale structure analysis and generation. 2020. arXiv:2006.06298.

[2] Mathieu Andreux, Tomás Angles, Georgios Exarchakis, Roberto Leonarduzzi, Gaspar Rochette, Louis Thiry, John Zarka, Stéphane Mallat, Joakim Andén, Eugene Belilovsky, Joan Bruna, Vincent Lostanlen, Muawiz Chaudhary, Matthew J. Hirn, Edouard Oyallon, Sixin Zhang, Carmine Cella, and Michael Eickenberg. Kymatio: Scattering transforms in python. *Journal of Machine Learning Research*, 21(60):1–6, 2020.

[15] We found that the bottleneck distance also suggested in [11] is not sufficiently discriminating for our point patterns.

Figure 9: TDA. Every point of a given cloud represents one entire realization of points in W_s sampled from the original distribution or using our generative model. Relative positions of points in the clouds reflect the similiarties or differences between these realizations as captured by the PDs. The point corresponding to the original sample $\bar{\phi}$ being used to generate all the syntheses in the given model is marked by a black dot.

[3] F. Baccelli and J. O. Woo. On the entropy and mutual information of point processes. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 695–699, 2016.

[4] François Baccelli, Bartłomiej Błaszczyszyn, and Mohamed Karray. *Random Measures, Point Processes, and Stochastic Geometry*. 2020.

[5] E. Bacry, A. Kozhemyak, and J. F. Muzy. Continuous cascade models for asset returns.

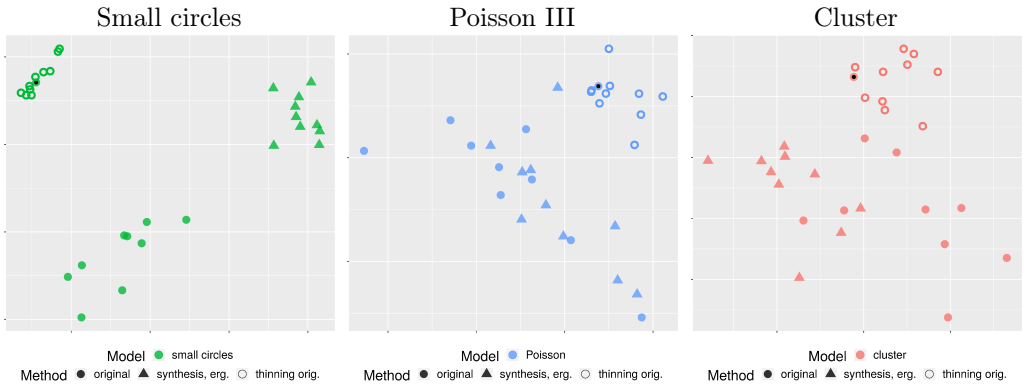

Figure 10: Effect of thinning (random choosing of 2000 points) on TDA. Ten dots labeled "thinning orig." (empty circles) represent point patterns which are obtained simply by independent thinning of the single realisation used for the synthesis (marked by the black dot).

Journal of Economic Dynamics and Control, 32:156–199, 2008.

[6] Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and topological inference*, volume 57. Cambridge University Press, 2018.

[7] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[8] Nicolas Bonnotte. Unidimensional and evolution methods for optimal transportation, 2013.

[9] Pierre Brémaud. *Mathematical principles of signal processing: Fourier and wavelet analysis*. Springer Science & Business Media, 2013.

[10] Joan Bruna and Stephane Mallat. Multiscale Sparse Microcanonical Models. *Math. Stat. Learn.*, 1(5):257–315, 2018.

[11] F. Chazal and B. Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. 2017. arXiv:1710.04019.

[12] Sung Nok Chiu, Dietrich Stoyan, Wilfrid S Kendall, and Joseph Mecke. *Stochastic Geometry and its Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK, aug 2013.

[13] D J Daley and D Vere-Jones. *An Introduction to the Theory of Point Processes, volume I: Elementary Theory and Methods of Probability and its Applications*. Springer, New York, 2003.

[14] David Dereudre. *Introduction to the Theory of Gibbs Point Processes*, pages 181–229. 04 2019.

[15] Lauris Ducasse and Alain Pumir. Intermittent particle distribution in synthetic free-surface turbulent flows. *Phys. Rev. E*, 77:066304, Jun 2008.

[16] Lauris Ducasse and Alain Pumir. Inertial particle collisions in turbulent synthetic flows: Quantifying the sling effect. *Phys. Rev. E*, 80:066312, Dec 2009.

[17] Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, and Clément Maria. Introduction to the R package TDA. *arXiv preprint arXiv:1411.1830*, 2014.

[18] L. A. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, pages 262–270, 2015.

[19] U Greb and M G Rusbridge. The interpretation of the bispectrum and bicoherence for non-linear interactions of continuous spectra. *Plasma Physics and Controlled Fusion*, 30(5):537–549, may 1988.

[20] Byrd R. H., Lu P., and Nocedal J. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 116 (5):1190–1208, 1995.

[21] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

[22] R. Leonarduzzi, G. Rochette, J.-P. Bouchaud, and S. Mallat. Maximum-entropy scattering models for financial time series. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

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

[24] Stéphane Mallat, Sixin Zhang, and Gaspar Rochette. Phase harmonic correlations and convolutional neural networks. *Information and Inference: A Journal of the IMA*, 11 2019.

[25] K. Matsuda and R. Onishi. Turbulent enhancement of radar reflectivity factor for polydisperse cloud droplets. *Atmos. Chem. Phys.*, 19:1785, 2019.

[26] Yves Meyer. *Wavelets and Operators*, volume 1 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 1993.

[27] Thibault Oujia, Keigo Matsuda, and Kai Schneider. Divergence and convergence of inertial particles in high-reynolds-number turbulence. *Journal of Fluid Mechanics*, 905:A14, 2020.

[28] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.

[29] Kai Schneider, Jörg Ziuber, Marie Farge, and Alexandre Azzalini. Coherent vortex extraction and simulation of 2d isotropic turbulence. *Journal of Turbulence*, 7(7):N44, 2006.

[30] R. S. Stoica, V. J. MartÃnez, J. Mateu, and E. Saar. Detection of cosmic filaments using the candy model. *Astronomy and Astropphysics*, 434(2):423–432, 2005.

[31] E. Tempel, R. S. Stoica, R. Kipper, and E. Saar. Bisous model - detecting filamentary patterns in point processes, 2016.

[32] S. Torquato. *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*. Springer, New York,, 1993.

[33] A. Tscheschel and D. Stoyan. Statistical reconstruction of random point patterns. *Computational Statistics and Data Analysis*, 51(2):859–871, nov 2006.

[34] Raoul R Wadhwa, Drew FK Williamson, Andrew Dhawan, and Jacob G Scott. TDAstats: R pipeline for computing persistent homology in topological data analysis. *Journal of open source software*, 3(28):860, 2018.

[35] T. Wiegand and K. A. Moloney. *Handbook of Spatial Point-Pattern Analysis in Ecology*. CRC Press, Taylor & Francis Group, 2014.

[36] Sixin Zhang and Stéphane Mallat. Maximum entropy models from phase harmonic covariances. *arXiv preprint arXiv:1911.10017*, 2019.

A Regularized particle gradient descent model

In this section we detail the regularized particle gradient descent model introduced in Section 4.2.2. This model is an alternative to the ergodic model, where the maximal scale parameter J is large ($J = O(\log(s))$). It consists in adding a regularization term to the objective function, in order to preserve the diversity of the model.

The regularization that we introduce in this section aims at preventing the successive iterations of the gradient descent model considered in Section 3.1 to move too far away from its initial configuration. It is achieved by adding the distance between the initial configuration and optimized configuration as a penalization term in the objective. It thus forces the model to explore local minima of $E_{\bar{\phi}}(\cdot)$ around the initial configuration. As all the configurations are measures with equal

mass, it is natural to consider the Wasserstein distance between them. However, as this distance is computationally expensive to compute, we choose to replace it with the sliced Wesserstein distance [7]. For counting measures with same mass, the 2-Wasserstein distance is defined as:

$$\mathcal{W}_2(\mu,\nu) = \min_{\sigma\in\mathfrak{S}_N} \Big(\sum_{i=1}^{N} ||x_i - y_{\sigma(i)}||^2\Big)^{\frac{1}{2}},$$

where the minimum is taken over the set $\mathfrak{S}_N$ of permutations of $(1,\dots,N)$.

The square $\mathcal{SW}_2$ distance between two planar counting measures μ and ν is then defined as follows:

$$\mathcal{SW}_2^2(\mu,\nu) := \int_{\theta\in\mathbb{S}^1} \mathcal{W}_2^2(\theta_{\#}\mu, \theta_{\#}\nu)d\theta,$$

where $\theta_{\#}$ designates the pushforward by the orthogonal projection operator in the direction θ.

The use of $\mathcal{SW}_2$ instead of $\mathcal{W}_2$ is motivated by the fact that applying the projection operators leads us to compute $\mathcal{W}_2$ on $\mathbb{R}$ instead of $\mathbb{R}^2$, which is much faster, as we simply need to sort the Dirac measures to obtain the optimal matching. It has been shown [8] that $\mathcal{SW}_2$ is indeed a distance, and that it induces the same topology as $\mathcal{W}_2$ for compact domains. One can approximate $\mathcal{SW}_2$ by choosing a certain number of fixed directions [7].

We use this distance with respect to the initial configuration $\phi_0 \in \mathbb{M}$ as a regularization term in the minimization of $E_{\bar{\phi}}(\cdot)$ defined in Eq. (9). More precisely, we consider the following optimization problem

$$\arg\min_{f\in\mathcal{F}(\mathbb{R}^2)} \frac{1}{2}||K(f_{\#}\phi_0) - K(\bar{\phi})||_H^2 - \lambda\frac{1}{2}\mathcal{SW}_2^2(f_{\#}\phi_0,\phi_0), \tag{21}$$

where the minimization is done over the set $\mathcal{F}(\mathbb{R}^2)$ of measurable functions from $\mathbb{R}^2$ into itself, and $\lambda \geq 0$ is a regularization parameter.

This new optimization problem implies that we do not try to sample from the microcanonical set here, as we suppose that it does not contain enough realizations of Φ.

Note, in (21) we optimize the transport $f_{\#}$ of a given initial configuration ϕ_0 towards $\bar{\phi}$, by minimizing the value of the function

$$\begin{aligned} E_{\bar{\phi}}^{\phi_0}(f) &= E_{\bar{\phi}}(f_{\#}\phi_0) - \lambda\frac{1}{2}\mathcal{SW}_2^2(f_{\#}\phi_0,\phi_0) \\ &= \frac{1}{2}||K(f_{\#}\phi_0) - K(\bar{\phi})||_H^2 - \lambda\frac{1}{2}\mathcal{SW}_2^2(f_{\#}\phi_0,\phi_0) \end{aligned}$$

In order to solve this problem we propose the following modification of the previously considered gradient descent model. For $\phi_0 \in \mathbb{M}$, $x \in \phi_0$, and any $f \in \mathcal{F}(\mathbb{R}^2)$, we define the functions

$$\begin{array}{rccc} h_{f,x}^{\phi_0} & : & \mathbb{R}^2 & \longrightarrow & \mathbb{M} \\ & & y & \longmapsto & f_{\#}\mu - \delta_{f(x)} + \delta_{f(x)+y}, \end{array}$$

and

$$\begin{array}{rccc} \mathcal{SW}_2{}_{f,x}^{\phi_0} & : & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^+ \\ & & y & \longmapsto & \mathcal{SW}_2(h_{f,x}^{\phi_0}(y),\phi_0). \end{array}$$

For any $x \in \phi_0$, and any $f \in \mathcal{F}(\mathbb{R}^2)$, $\mathcal{SW}_2{}_{f,x}^{\phi_0}$ is differentiable, so we can define a function from $\mathbb{R}^2$ to $\mathbb{R}^2$

$$x \longmapsto \nabla_x\mathcal{SW}_2(f_{\#}\phi_0,\phi_0) := \begin{cases} J[\mathcal{SW}_2{}_{f,x}^{\phi_0}](0) & \text{if } x \in \mathrm{Supp}(f_{\#}\phi_0) \\ 0 & \text{otherwise.} \end{cases}$$

We can then, for any initial measure ϕ_0 and any measurable function f, define a function from $\mathbb{R}^2$ to $\mathbb{R}^2$

$$x \longmapsto \nabla_x E_{\bar{\phi}}^{\phi_0}(f) := \lambda\nabla_x\mathcal{SW}_2(f_{\#}\phi_0,\phi_0) + \nabla_x E_{\bar{\phi}}(f_{\#}\phi_0),$$

where $\nabla_x E_{\bar{\phi}}$ is defined in (12).

The sequence of point-configuration transports induced by gradient descent algorithm is defined recursively by taking an initial configuration ϕ_0, $f_0 := I_d$ (identity) and for $n \geq 0$

$$\begin{cases} f_{n+1} : x \longmapsto x - \gamma \nabla_x E^{\phi_0}_{\bar{\phi}}(f_n \circ \ldots \circ f_0) \\ \phi_{n+1} = f_{n+1\,\#}\phi_n = (f_n \circ \ldots \circ f_0)_{\#}\phi_0. \end{cases}, \tag{22}$$

where $\gamma > 0$ is some gradient step size.

In full analogy to (13) and (14), for all $n \geq 1$, ϕ_n transport the initial distribution $\mathcal{L}_0$ (of ϕ_0) to some $\mathcal{L}_n$. Note $\mathcal{L}_k = \mathcal{L}_{\Phi_n}$ is the distribution of the point process Φ_n obtained by n iterations of (22) started of Φ_0 having law $\mathcal{L}_0 = \mathcal{L}_{\Phi_0}$.

B Proof of Theorem 3.3

Proof of Theorem 3.3. We are going to show that for all $n \in \mathbb{N}$, $\mathcal{L}_{\bar{\Phi}_n}$ defined as the push-forward of $\mathcal{L}_{\bar{\Phi}_0}$ by (14) is invariant to the action of T. The proof also applies to the regularized model.

As the descriptor K is invariant to the action of T, we do not precise the dependency of the following functions to $\bar{\phi}$, even though $E_{\bar{\phi}}$ clearly depends on the observation $\bar{\phi}$. The result follows from the fact (which we shall prove later) that for each $n \geq 0$, the function $\mathbb{R}^2 \ni x \longmapsto f_n(x) = f_n(x, \phi_0)$ defined in (22) satisfies

$$f_n(Tx; T_{\#}\phi_0) = T f_n(x; \phi_0), \quad \phi_0 \in \mathbb{M}_{W_o} \tag{23}$$

and, by consequence, the same holds for the composition $F_n(x, \phi_0) := f_n \circ \ldots \circ f_1(x, \phi_0)$.

Indeed, if (23) holds, then for $\Phi_n := F_{n\,\#}\Phi_0 = F_n(\cdot, \Phi_0)_{\#}\Phi_0$ and $\Gamma \in \mathcal{M}_{W_0}$

$$\begin{aligned} \mathcal{L}_{\Phi_n}(T_{\#}^{-1}(\Gamma)) &= \mathcal{L}_{\Phi_0}\left\{\phi_0 : F_n(\cdot, \phi_0)_{\#}\phi_0 \in T_{\#}^{-1}(\Gamma)\right\} \\ (*) &= \mathcal{L}_{\Phi_0}\left\{\phi_0 : F_n(\cdot, T_{\#}^{-1}\phi_0)_{\#}T_{\#}^{-1}\phi_0 \in T_{\#}^{-1}(\Gamma)\right\} \\ &= \mathcal{L}_{\Phi_0}\left\{\phi_0 : T_{\#}F_n(\cdot, T_{\#}^{-1}\phi_0)_{\#}T_{\#}^{-1}\phi_0 \in \Gamma\right\} \\ &= \mathcal{L}_{\Phi_0}\left\{\phi_0 : T \circ F_n(\cdot, T_{\#}^{-1}\phi_0)_{\#}T_{\#}^{-1}\phi_0 \in \Gamma\right\} \\ (**) &= \mathcal{L}_{\Phi_0}\left\{\phi_0 : F_n(T\cdot, \phi_0)_{\#}T_{\#}^{-1}\phi_0 \in \Gamma\right\} \\ &= \mathcal{L}_{\Phi_0}\left\{\phi_0 : F_n(T \circ T^{-1}\cdot, \phi_0)_{\#}\phi_0 \in \Gamma\right\} \\ &= \mathcal{L}_{\Phi_0}\left\{\phi_0 : F_n(\cdot, \phi_0)_{\#}\phi_0 \in \Gamma\right\} \\ &= \mathcal{L}_{\Phi_n}(\Gamma), \end{aligned}$$

where the equality $(*)$ follows from the above equality (23). and the equality $(**)$ follows from (23) applied to the composition with F_n. It remains to prove (23).

Proof of (23) *follows by induction.*
Note $f_0 = I_d$ trivially satisfies the equality. Assume now that (23) is satisfied for all f_k, $k = 1, \ldots, n$ and, consequently, for $F_n = f_n \circ \ldots \circ f_1$. It remains to prove that the function

$$f_{n+1}(x, \phi_0) := x - \nabla_x \mathcal{SW}_2(F_n(\cdot, \phi_0)_{\#}\phi_0, \phi_0) - \nabla_x E_{\bar{\phi}}(F_n(\cdot, \phi_0, \phi)_{\#}\phi_0) \tag{24}$$

is equivariant as in (23) (where, for the simplicity of the proof, and without loss of generality, we have assumed $\lambda = \gamma = 1$).

Proof of (23).
We will frequently use the following relation resulting from the induction assumption

$$F_n(\cdot, T_{\#}\phi_0)_{\#}T_{\#}\phi_0 = F_n(T\cdot, T_{\#}\phi_0)_{\#}\phi_0 = T F_n(\cdot, \phi_0)_{\#}\phi_0. \tag{25}$$

Case $x \notin F_n(\cdot, \phi_0)_{\#}\phi_0$.
In this case $f_{n+1}(x, \phi_0) = x$ since, by the definition, both gradients are null. Also, by (25), ad

because T is invertible, $Tx \notin TF_n(\cdot,\phi_0)_\#\phi_0 = F_n(\cdot,T_\#\phi_0)_\#T_\#\phi_0$ and hence $f_{n+1}(Tx,T_\#\phi_0) = Tx$, which proves (24) for the considered x.

Case $x = F_n(x_i,\phi_0)$ for some $x_i \in \phi_0$.

In what follows, we shall simplify the notation replacing $F_n(\cdot,\phi_0)_\#\phi_0$ by F_n and $F_n(x_i,\phi_0)$ by $F_n(x_i)$.

The required equality (24) in this case follows from the following two gradient relations which we prove later

$$\nabla_{TF_n(x_i)} E_{\bar\phi}(TF_n) = A\nabla_{F_n(x_i)} E_{\bar\phi}(F_n) \tag{26}$$

and

$$\nabla_{TF_n(x_i)} \mathcal{SW}_2(TF_n, T_\#\phi_0) = A\nabla_{F_n(x_i)} \mathcal{SW}_2(F_n,\phi_0)\,. \tag{27}$$

Indeed, applying (25) and both relations (26), (27) to $f_{n+1}(x,\phi_0)$ in (25) we have

$$\begin{aligned} f_{n+1}(Tx_i,T_\#\phi_0) &= Tx_i - \nabla_{TF_n(x_i)}\mathcal{SW}_2(TF_n,T_\#\phi_0) \\ &\quad - \nabla_{TF_n(x_i)} E_{\bar\phi}(TF_n) \\ &= Tx_i - A\nabla_{F_n(x_i)}\mathcal{SW}_2(F_n,\phi_0) \\ &\quad - A\nabla_{F_n(x_i)} E_{\bar\phi}(F_n) \\ (***) &= T\Big(x_i - \nabla_{F_n(x_i)}\mathcal{SW}_2(F_n,\phi_0) \\ &\quad - \nabla_{F_n(x_i)} E_{\bar\phi}(F_n)\Big) \\ &= Tf_{n+1}(x_i,\phi_0)\,, \end{aligned}$$

where in $(***)$ we have used again $Tx - Ay - Az = T(x-y-z)$. It remains to prove (26) and (27).

Proof of (26) *and* (27).

We have

$$\begin{aligned} &\nabla_{TF_n(x_i)} E_{\bar\phi}(TF_n) \\ &= \Big(\nabla_{TF_n(x_i)} K(TF_n)\Big)^t \Big(K(TF_n) - K(T_\#\phi)\Big) \end{aligned} \tag{28}$$

and, using the fact that $A^{-1} = A^t$,

$$\begin{aligned} \nabla_{TF_n(x_i)} K(TF_n) &= J[K^{TF_n}_{TF_n(x_i)}](0) \\ &= J[K \circ h^{TF_n}_{TF_n(x_i)} \circ A](0)A^t \end{aligned} \tag{29}$$

with

$$\begin{aligned} &K \circ h^{TF_n}_{TF_n(x_i)} \circ A(y) \\ &= K\Big(TF_n - \delta_{TF_n(x_i)} + \delta_{TF_n(x_i)+Ay}\Big) \\ &= K\Big(TF_n - \delta_{TF_n(x_i)} + \delta_{T(F_n(x_i)+y)}\Big) \\ &= K\Big(T_\#\big(F_n - \delta_{F_n(x_i)} + \delta_{F_n(x_i)+y}\big)\Big) \\ &= K\Big(F_n - \delta_{F_n(x_i)} + \delta_{F_n(x_i)+y}\Big) \\ &= K \circ h^{F_n}_{F_n(x_i)} \end{aligned} \tag{30}$$

where the last but one equality follows from the invariance of K with respect to T. Using (29) and (30) in (28) we obtain

$$\begin{aligned} &\nabla_{TF_n(x_i)} E_{\bar\phi}(TF_n) \\ &= \Big(\nabla_{F_n(x_i)} K(F_n)A^t\Big)^t \Big(K(F_n - (T_\#\phi)\Big) \\ &= A\Big(\nabla_{F_n(x_i)} K(F_n\Big)^t \Big(K(F_n - (T_\#\phi)\Big) \\ &= A\nabla_{F_n(x_i)} E_{\bar\phi}(F_n). \end{aligned}$$

which proves (26). The proof of (27) follows the same lines, as the gradient of the regularization function is equivariant to T and the regularization itself is invariant. This concludes the proof of the equivariance of f_{n+1} in (24) and thus of Theorem 3.3. □

C Fourier analysis and its comparison to TDA

In this section we briefly present descriptors based on Fourier (or power) spectrum of point measures and remind the relations to Bartlett spectrum of point processes. We also compare spectrum analysis to TDA.

C.1 Fourier spectrum based descriptors

We define the *discrete Fourier transform* (DFT) $F_m(\mu)$ of a counting measure $\mu = \sum_u \delta_{x_u} \in \mathbb{M}^s$ on the (square) window $[-s, s[^2$ at integer frequency $m \in \mathbb{Z}^2$ by

$$F_m(\mu) := \int_{W_s} e^{-i\pi m x/s}\, \mu(dx) = \sum_u e^{-i\pi m x_u/s}.$$

Observe, $F_m(\mu)$ at frequency $m = (0,0)$ specifies the number of points of the measure μ on W_s. *Fourier spectrum* (or *power spectrum*) (PS) is often defined by taking the square modulus of the Fourier coefficients $F_m(\mu)$; $U_m(\mu) := |F_m(\mu)|^2/|W_s|$, normalized for convenience by the window size. Note $|F_m(\mu)|^2$, and consequently $U_m(\mu)$ is invariant with respect to (circular) translations of μ on W_s. By selecting the frequencies in a limited range $m \in \Gamma_F \subset \mathbb{Z}^2$ one obtains a translation-invariant Fourier spectrum (or PS) based descriptor. As we shall focus on isotropic point processes, we further reduce the variance of our descriptor by averaging Fourier coefficients along frequency orientations. More precisely, let us define $\tilde{\Gamma}_F := \{\lfloor |m| \rfloor,\ m \in \Gamma_F\}$. For each $k \in \tilde{\Gamma}_F$, we define $\tilde{U}_k(\mu) := \frac{1}{\#k} \sum_{\substack{m \in \Gamma_F \\ \lfloor |m| \rfloor = k}} U_m(\mu)$, where $\#k$ denotes the cardinal of $\{m \in \Gamma_F : \lfloor |m| \rfloor = k\}$. We define our rotation and translation invariant descriptor:

$$K_F(\mu) = \{\tilde{U}_k(\mu)\}_{k \in \tilde{\Gamma}_F}.$$

This descriptor captures only information about 2nd order correlations between points of μ, which is *not* enough for satisfactory reconstruction (and hence synthesis) of point patterns. To illustrate the above statement, we define a particle gradient descent model over $\mathbb{M}^s$ using K_F as follows. To transport the measure μ in the continuous domain W_s efficiently, we choose a normalized Euclidean metric for the logarithm of the descriptors K_F in H which makes the optimization problem well conditioned. It amounts to minimize the energy

$$E^F_{\bar{\phi}}(\mu) := \frac{1}{2} \sum_{k \in \tilde{\Gamma}_F} \left(\log(|\tilde{U}_k(\mu)|/|\tilde{U}_k(\bar{\phi})|)\right)^2. \tag{31}$$

(In case there are zeros in $\{\tilde{U}_k(\mu)\}_{k \in \tilde{\Gamma}_F}$, we may add a small constant $\epsilon > 0$ to both $|\tilde{U}_k(\mu)|^2$ and $|\tilde{U}_k(\phi)|^2$ in $E_{\bar{\phi}}(\mu)$.) This problem can be solved numerically, as the gradient of $|\tilde{U}_k(\mu)|^2$ with respect to each point in μ can be computed analytically. With this descriptor, contrary to wavelet phase harmonics descriptors (eq. (19)), the procedures of Section 5 are not necessary: the DFT can be computed directly from the positions of the atoms, so no discretization is needed. This implies that no final blurring is needed either, as it is used to overcome discretization issues. Furthermore, we empirically observe that no multiscale procedure is needed. In Figure 11, we evaluate the particle gradient descent model with this energy with $\Gamma_F = [-128, 128[^2$. We see that the Fourier descriptors are well matched between the original image, which is the Turbulent Poisson II model presented on Figure 6, and the samples generated using this Fourier-based descriptor. However, visually, the geometric structure exhibited by the the generated sample is different form the the original configuration.

One way to explain this result is that $F_m(\bar{\Phi})$ and $F_{m'}(\bar{\Phi})$ may still have strong dependencies which are not respected by the samples $\mu \in \Omega_\epsilon$. To capture their dependencies, one may consider

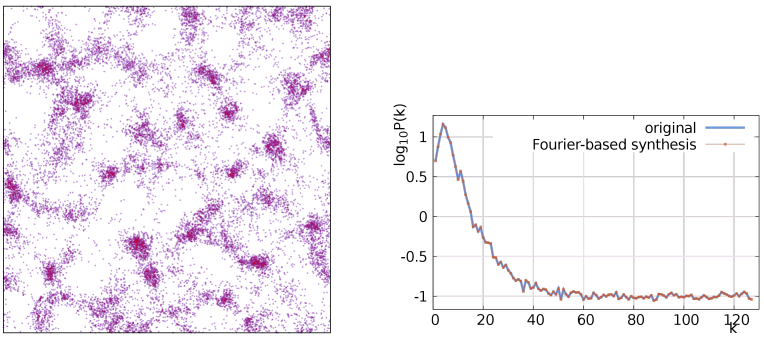

Figure 11: Left: An attempt to reconstruct a Turbulent Poisson II configuration (see Figure 6 for the original configuration) using Fourier-based energy (31) with $\Gamma_F = [-128, 128[^2$. Right: (Rotationally averaged) power spectrum plot of the generated sample exactly matching the original.

using high-order statistics. For example, the bi-spectrum is also a translation invariant descriptor ([19]) of the form

$$F_m(\mu)F_{m'}(\mu)F_{m+m'}(\mu)^*$$

They capture partially the dependencies between the Fourier coefficients at $(m, m') \in \Gamma_F^2$. A major issue with descriptors based on high-order statistics is their empirical variance, which measures the sensitivity of a model to outliers in $\bar{\phi}$. That is why we use the wavelet transform and their phase harmonics to capture dependencies across frequencies without using higher-order statistics.

C.2 Bartlett spectrum of point processes

We now briefly recall the relations of the PS to the Bartlett spectrum of point processes. To this end, assume that $\bar{\Phi}$ is a realization of a stationary point process Φ truncated to the observation window W_s. We have $\mathbb{E}[F_{(0,0)}(\bar{\Phi})] = \mathbb{E}[\Phi(W_s)]$. However, for $m \in \mathbb{Z}^2 \setminus \{(0,0)\}$ $\mathbb{E}[F_m(\bar{\Phi})] = 0$. (This follows, by the Campbell formula, from the fact that the mean measure of a stationary point process — which is a constant multiple of the Lebesgue measure — integrates $e^{-i\pi mx/s}$ over the square window W_s to zero.) Consequently, by the definition of the *Bartlett spectral measure* B_Φ of Φ (cf. [9, Definition 5.2.3])

$$\begin{aligned}\mathbb{E}[U_m(\bar{\Phi})] = \mathbb{E}[|F_m(\bar{\Phi})|^2/|W_s|] &= \mathrm{Var}(F_m(\bar{\Phi}))/|W_s| \\ &= \frac{1}{4s^2}\int_{\mathbb{R}^2} |\widehat{1_{W_s}}(\xi - m/(2s))|^2 \, B_\Phi(d\xi) \\ &= \frac{1}{4s^2}\int_{\mathbb{R}^2} G_s^2(\xi - m/(2s)) \, B_\Phi(d\xi),\end{aligned}$$

where $\widehat{f}(\xi) = \int_{\mathbb{R}^2} f(x)e^{-2i\pi\xi x}\,dx$ is the Fourier transform of f, 1_{W_s} is the indicator function of the square (Dirichlet) window W_s, and $G_s(\xi) = 4s^2\,\mathrm{sinc}(2s\xi_1)\mathrm{sinc}(2s\xi_2)$ with $\xi = (\xi_1, \xi_2)$, $\mathrm{sinc}(t) = \sin(\pi t)/(\pi t)$ its Fourier transform; cf. [9, Example 1.1]. In words, the expected spectral power $\mathbb{E}[U_m(\bar{\Phi})]$ equals to the convolution of the Bartlett spectral measure of Φ with the window function $G_s^2/(4s^2)$ evaluated at the frequency $m/2s$. Note the function $G_s/(4s^2)$ integrates to 1 and for s large enough and B_Φ admitting continuous density b_Φ

$$\mathbb{E}[U_m(\bar{\Phi})] \approx b_\Phi(m/(2s)). \tag{32}$$

We can observe that the above approximation is quite accurate for the circle-Cox point process considered in Section 6, where b_Φ can be evaluated explicitly. Indeed, the density of the Bartlett spectrum is related via the Fourier transform

$$b_\Phi = \lambda + \lambda^2(\widehat{g_\Phi - 1})$$

to the pair correlation function g_Φ of Φ. This latter function can be calculated for the Boolean circle model with Poisson points of linear intensity λ' on circles of radius R

$$g_{\mathrm{BC}}(x) = 1 + \begin{cases} \frac{2\lambda'}{\pi|x|\lambda} \frac{R}{\sqrt{4R^2-|x|^2}} & \text{for } |x| \leq R \\ 0 & \text{otherwise,} \end{cases}$$

where $x = (x_1, x_2) \in \mathbb{R}^2$, $|x| = \sqrt{x_1^2 + x_2^2}$, (cf. [12, Example 10.6]) yielding

$$b_{\mathrm{BC}(}\xi) = \lambda(1 + 2\pi R\lambda' J_0^2(2\pi R|\xi|)), \tag{33}$$

where J_0 is the 0-th Bessel function of the first kind.

C.3 Spectrum evaluation vs TDA

In this section we illustrate some differences between TDA and power spectrum evaluation, by showing cases where one tool has strong discriminate power while the other does not. Furthermore, this comparison illustrates the consistence between visual appreciation and TDA.

First, we consider, for the Small circles Cox process, two different models: the one we presented in Section 6.2, and a model where both DFT and our 'ergodic' descriptor are used. As we can see in Figure 12, while the power spectrum of the two models are quite different, one cannot discriminate them visually, and the TDA visualization of both models are close (compared to the original distribution).

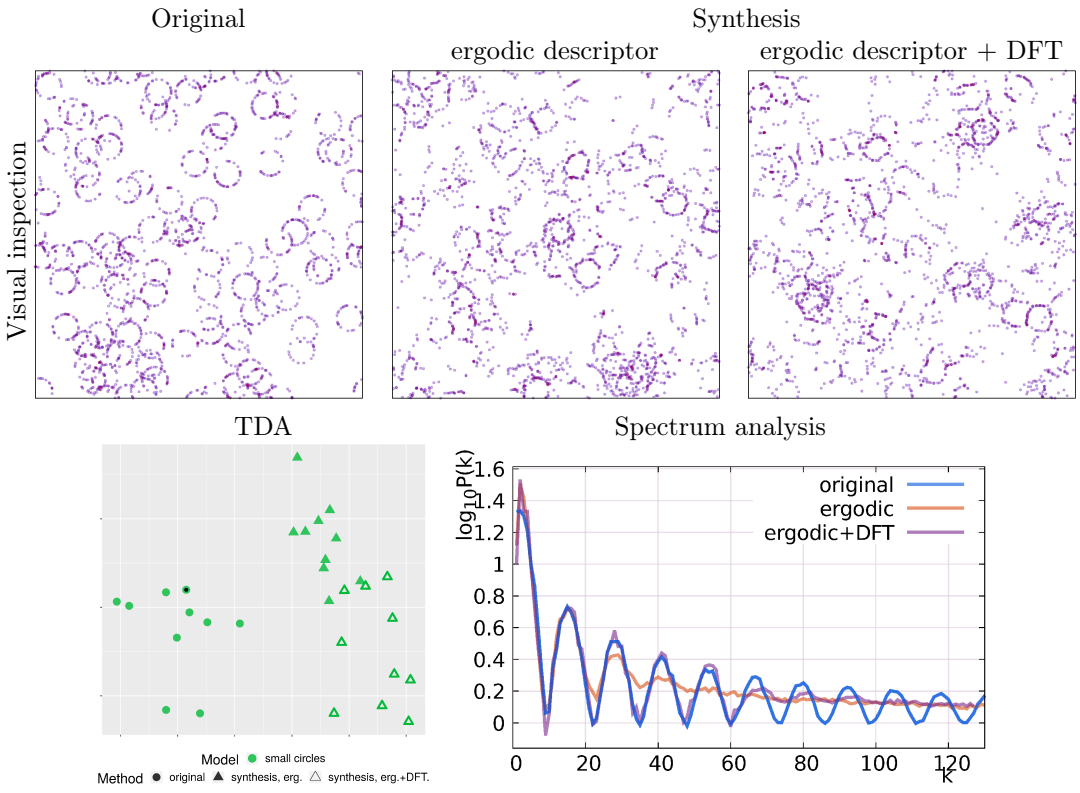

Figure 12: Synthesis of small circles model: visual inspection, TDA and the spectrum analysis. Adding DFT to the (ergodic) descriptor does not improve visual perception despite the fact that it significantly improves the spectrum matching. TDA seems to corroborate with the visual inspection showing small distance between patterns generated with the two descriptors.

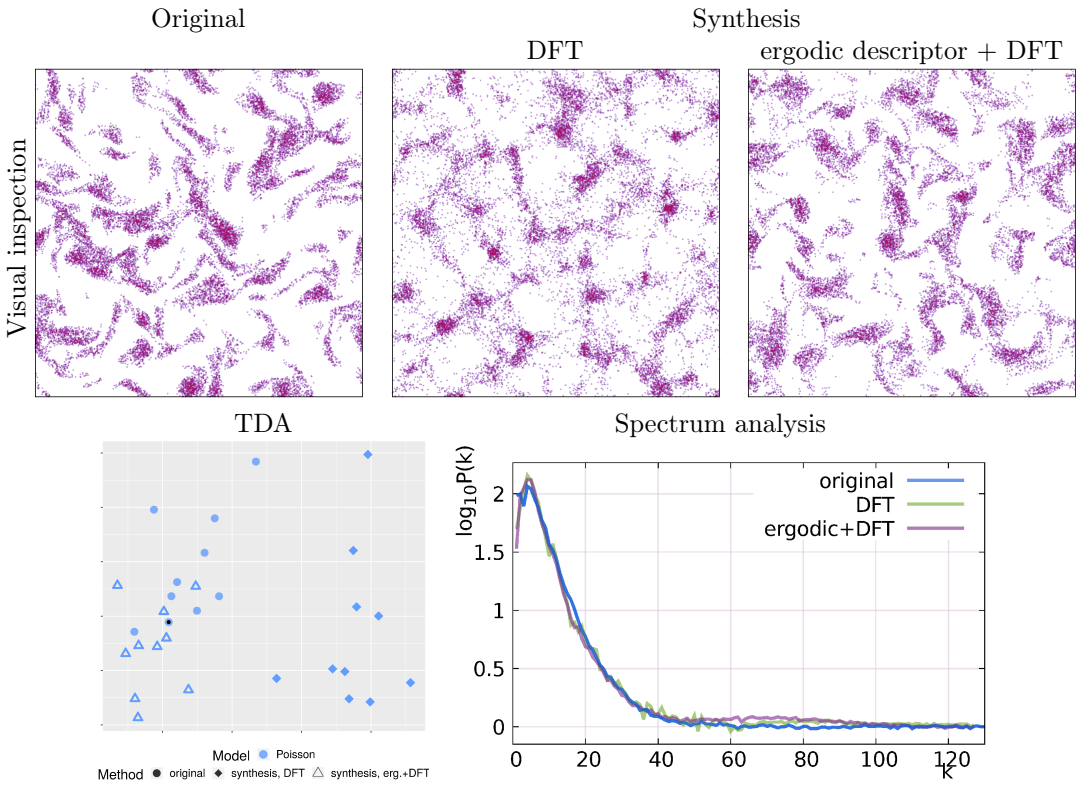

Figure 13: Synthesis of the turbulent Poisson II models using DFT moments and DFT+ergodic moments: visual inspection, TDA and the spectrum analysis. There is no big difference between the spectrum for the two synthesis methods (the spectrum of all 10 synthesis with DFT perfectly matches the spectrum of the original pattern, cf. Figure 11, and this is why there is no averaging effect of the corresponding plot). A clear visual difference is confrmed by the TDA.