Particle gradient descent model for point process generation

Antoine Brochard, Bartlomiej Blaszczyszyn, Sixin Zhang, Stéphane Mallat

Noname manuscript No.
(will be inserted by the editor)

Particle gradient descent model for point process generation

Antoine Brochard · **Bartłomiej Błaszczyszyn** · **Sixin Zhang** · **Stéphane Mallat**

the date of receipt and acceptance should be inserted later

Abstract This paper presents a statistical model for stationary ergodic point processes, estimated from a single realization observed in a square window. With existing approaches in stochastic geometry, it is very difficult to model processes with complex geometries formed by a large number of particles. Inspired by recent works on gradient descent algorithms for sampling maximum-entropy models, we describe a model that allows for fast sampling of new configurations reproducing the statistics of the given observation. Starting from an initial random configuration, its particles are moved according to the gradient of an energy, in order to match a set of prescribed moments (functionals). Our moments are defined via a phase harmonic operator on the wavelet transform of point patterns. They allow one to capture multi-scale interactions between the particles, while controlling explicitly the number of moments by the scales of the structures to model. We present numerical experiments on point processes with various geometric structures, and assess the quality of the model by spectral and topological data analysis.

Keywords Point processes, Simulation model, Entropy, Wavelets, Spectral analysis, Topological data analysis

A. Brochard
Inria/ENS/PSL University, Paris, France
E-mail: antoine.brochard@inria.fr
Huawei Technologies France, Boulogne-Billancourt, France

B. Błaszczyszyn
Inria/ENS/PSL University, Paris, France

S. Zhang
IRIT, Université de Toulouse, CNRS, Toulouse, France

S. Mallat
ENS/PSL University, Paris, France
Collège de France, Paris, France
Flatiron Institute, New York, USA

Declarations

Fundings : This work was partly supported by the PRAIRIE 3IA Institute of the French ANR-19-P3IA-0001 program. Sixin Zhang was 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.

Conflict of interest The authors declare that they have no conflict of interest.

Availability of data and material For the sake of transparency, we are ready to make available the data used to produce the results in our paper.

Code availability For the sake of transparency, we are ready to make available the code used to produce the results in our paper.

1 Introduction

In order to generate new realizations of a stochastic process of which we have only one realization, we have to build a probabilistic model which approximates the distribution of this process, and from which we can sample. In this article, we are interested in generative models for stationary, ergodic point processes. Such models are of interest in a wide range of applications (Illian et al. 2008, Chapter 6), for instance biology (Diggle et al. 2006; Baddeley et al. 2014), ecology (Wiegand and Moloney 2013), turbulent flows in atmosphere science (Ducasse and Pumir 2008, 2009; Matsuda and Onishi 2019; Oujia et al. 2020), or cosmology (Stoica et al. 2005; Tempel et al. 2016). In some of these domains, the observed patterns exhibit complex structures, with a large number of

particles (such as filaments in cosmology, or vortexes in turbulent flows). Our work is motivated by the simulation of such processes.

In this paper, we seek to generate realizations formed by a large number of particles, with both short and long range interactions. Figure 1 shows some examples of distributions that we shall consider. Currently, for such complex and diverse geometries, which naturally appear e.g. in cosmology or turbulent flows in physics and atmosphere science, no model has been proposed in the literature on point processes. To address this problem, we shall introduce a statistical model developed from the maximum-entropy principle (Jaynes 1957), to approximate such point process distributions and simulate new realizations.

Maximum-entropy models are based on the description of the distribution with a set of moments. Intuitively, this means that the model is 'as random as possible' under certain constraints, based on the information captured by the moments. There are three underlying problems in defining such models:

1. Choosing the moments that will describe the distribution. They should be informative enough to capture the geometric structures characterizing the distribution. On the other hand, they should be accurately estimated from a single observation, so the number of moments should not be too large.
2. Specifying a model deriving from these moments. This can be done by defining a maximum entropy model such as the macro-canonical model (maximizing the entropy under expectation constraints), or the micro-canonical model (maximizing the entropy under path-wise constraints).
3. Generating new samples from the model. In the micro-canonical setup, this can be done by minimizing an energy, that defines the set of admissible realizations. The minimization method must make it possible to generate diverse low energy samples without being too costly in terms of calculation.

In this paper, we shall place our model in the micro-canonical setup, detailed in Section 2. The main challenges reside in the problems 1 and 3. In this regard, we present multi-scale moments, new in the literature on point process, as well as a fast sampling algorithm based on gradient descent.

In Section 3, we present our method to address the problem of generating new samples: we minimize the energy of a new sample by moving the particles of an initial random configuration using the gradient of its energy with respect to the particles positions. In the point process literature, a classical method (Tscheschel and Stoyan 2006) consists in updating an initial random configuration by successively replacing the particles one by one, with new particles located at random positions (we shall call this method random search in this paper). The major drawback of this method is its computational cost, as the optimization, which does not use gradient information to minimize the energy, requires a large number of energy evaluations. In fact, this method has been applied to generate point processes formed by a few hundred particles. On the other hand, advanced methods in the modelling of textures and non-Gaussian stationary processes allow for fast sampling by first drawing from an initial Gaussian distribution, and minimizing an energy by gradient descent on the amplitudes of the pixels of the image (Portilla and Simoncelli 2000; Gatys et al. 2015; Bruna and Mallat 2019; Zhang and Mallat 2021). Our approach leverages the efficiency of this sampling method, while ensuring that the resulting samples are atomic measures. The idea of moving the points according to their gradient is often used in molecular dynamics (Zhang et al. 2015a,b), however it requires knowledge of the physical mechanisms behind the underlying process. Our statistical modeling approach has a potential to simulate new, complex particle configurations directly from one observation, when the underlying physical phenomena are very complicated to model.

This brings us to the other challenge that we address in this work: choosing the moments that we shall use to characterize the distribution. In Section 4, we present the wavelet phase harmonic covariance moments for point processes. These are spatial statistics based on coefficients computed from a wavelet transform of atomic measures, i.e. the convolution of the atomic measures with continuous local functions. It is known that the covariance between the wavelet coefficients capture only second-order correlations (Brémaud 2002, Section 5.2), which are equivalent to the Bartlett spectrum (Bartlett 1964). To capture information beyond second-order correlations, we apply a non-linear phase harmonic operator to the wavelet coefficients. This operator acts on the complex phase of the wavelet coefficients, without changing their amplitude (Mallat et al. 2020). The covariance between the resulting coefficients allows one to capture particles interactions across different scales. Compared to high-order correlation functions (Torquato 2002, Section 12.4.2), our moments have the potential to define a sufficient set of statistics, while maintaining a small estimation error, which is similar to the second-order statistics. Other statistics often used in the point process literature (e.g. the k nearest neighbour distribution function suggested in Tscheschel and Stoyan (2006)) have a number of elements that grows with the intensity. Since there is only one observation, the number of moments should be limited, in order to control their estimation variance. The wavelet transform allows for direct control over the scales of the structures that we wish to capture, regardless of the intensity of the process. This property allows one to model point processes formed by a large number of particles with a limited number of moments.

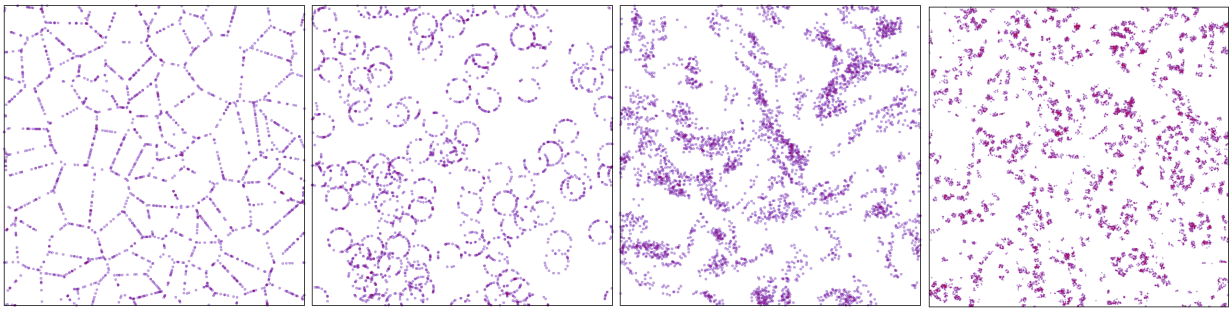

Fig. 1 Samples of point processes of various geometries. The number of points ranges from 1000-13000.

The wavelet phase harmonic covariance descriptors are defined as spatial averages evaluated over a point process realization (Zhang and Mallat 2021). In practice, the calculation of such descriptors can be done by discretization of the observation window in the form of a grid of pixels. However, making these descriptors differentiable with respect to the positions of the particles remains a challenge. In this regard, we describe in Section 5 a complete numerical scheme allowing one to solve this problem. It is based on a differentiable discretization of atomic measures. We further present a multiscale optimization in the gradient descent, intended to avoid unwanted shallow minima of the energy.

In Section 6, we evaluate our model on some distributions exhibiting various geometric structures, like Cox point processes on the edges of Poisson-Voronoi tessellations and on the Boolean model with circular grains. Other processes we consider are Matern hard-core and cluster processes driven by Poisson processes with turbulent intensities. Their intensities are sampled from a turbulent flow simulated from Navier-Stokes equations (Schneider et al. 2006). Besides the visual inspection of the samples from our generative model, we evaluate second order correlations and compare the persistent homology diagrams, that has been proven useful for topological data analysis (see e.g. Chazal and Michel (2017)).

In Section 7, we numerically compare our method with the classical approach developed in Tscheschel and Stoyan (2006) in terms of the speed of simulation, and the quality and diversity of the syntheses. Besides using the evaluation methodologies in Section 6, we also use a statistical moment matching approach suggested in Illian et al. (2008, Chapter 6). All the results can be reproduced by a software which is available at `https://github.com/abrochar/pp_syn`. A longer version of this paper is available, see Brochard et al. (2020).

Notations: For any integer $n \geq 1$ and any $z \in \mathbb{C}^n$, we note $|z|$ the Euclidean norm of z, and z^* is its complex conjugate. Let $\mathrm{Cov}(A,B) = \mathbb{E}[AB^*] - \mathbb{E}[A]E[B^*]$ denote the covariance between two complex random variables A and B. Let $\langle a,b\rangle$ denote the Euclidean inner product between two vectors $a \in \mathbb{R}^2$ and $b \in \mathbb{R}^2$.

2 Point process framework

In Section 2.1, we define the elementary objects of point process theory and the notations that we will use in this paper. A more detailed introduction to point processes and stochastic geometry can be found e.g. in Daley and Vere-Jones (2008); Chiu et al. (2013). We then review, in Section 2.2, the classical maximum-entropy models for point processes.

These models are theoretically well founded, but hard to sample from in general. Our model, presented in Section 3, takes inspiration from these, while being amenable to fast sampling.

2.1 General definitions

Configurations of points (on the plane) are represented as counting measures on $(\mathbb{R}^2, \mathscr{B})$, with $\mathscr{B}$ 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 $(\mathbb{R}^2, \mathscr{B})$, endowed with the σ-algebra $\mathscr{M}$ generated by the mappings $\mu \mapsto \mu(B)$, for $B \in \mathscr{B}$. 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

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

As a special case, for $x \in \mathbb{R}^2$, we define the translation $S_x \mu$ of μ by x, i.e. $S_x \mu(B) := \mu(B + x)$.

A 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 $\mu(\{x\}) > 0$.

A point process Φ is a measurable mapping from an abstract probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to $(\mathbb{M}, \mathscr{M})$. We will denote by $\mathscr{L}_\Phi$ the distribution of Φ, that is the pushforward of the probability measure $\mathbb{P}$ by Φ on $(\mathbb{M}, \mathscr{M})$. We say that a point process Φ is simple if $\mathbb{P}(\Phi \text{ is a simple measure}) = 1$. In this paper, for simplicity we shall only consider simple point processes.

Point process Φ is called *stationary* if its distribution $\mathscr{L}_\Phi$ is invariant with respect to all shifts S_x, $x \in \mathbb{R}^2$. It is said to be *ergodic* if the empirical averages (of real, measurable functions f on $\mathbb{M}$, integrable with respect to $\mathscr{L}_\Phi$) over windows $W_s = [-s,s[\times[-s,s[$ increasing to $\mathbb{R}^2$ converge almost surely to the mathematical expectations

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

where $|W_s|$ stands for the Lebesgue measure of W_s, see Daley and Vere-Jones (2008, Chapter 12) for more details.

For a given $s > 0$, we denote by $\mathbb{M}^s$ the set of counting measures on W_s, and $\mathscr{M}^s$ its induced σ-algebra. 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 $\mathbb{M}^s$ with torus correction on the window W_s.

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 remainder of this paper, we shall consider realizations of point processes observed on a finite square window $W_s = [-s,s]^2$, for some $s > 0$. We denote by $\bar{\Phi}$ the restriction of Φ to W_s, that is $\bar{\Phi}$ is a point process on W_s such that, $\forall n \in \mathbb{N}, \forall (B_1, ..., B_n) \in \mathscr{B}(W_s)^n$, $(\Phi(B_1), ..., \Phi(B_n)) = (\bar{\Phi}(B_1), ..., \bar{\Phi}(B_n))$ in distribution (where $\mathscr{B}(W_s)$ 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 Maximum entropy models for point processes

Maximum entropy models are based on the following intuitive idea: given an observation pattern, we aim at finding new patterns that are similar to, but different from the observation. To this end, we define a notion of similarity by choosing a set of statistics that will be computed on the observation and on the new patterns. The two will be considered similar if their statistics match. Furthermore, if the chosen statistics describe sufficiently well the point process behind our observation, we do not want to add any more constraints, that is, we want to find new patterns 'as random as possible', under the constraints defined by the statistics. This can be formalized by maximizing the entropy of the model.

This section defines both macro-canonical and micro-canonical models for a point process $\bar{\Phi}$ observed in the square window W_s. These models rely on maximizing the entropy a probability distribution under a set of moment constraints. They are used in large classes of stochastic models (Geman and Geman 1984), and will inspire our particle gradient descent model.

2.2.1 Point process entropy

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 (Baccelli and Woo 2016), 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 (Dereudre 2019). More specifically, let us denote by $\mathscr{L}_0$ the Poisson distribution on W_s. We define the KL divergence of a point process $\bar{\Phi}$ on W_s with distribution $\mathscr{L}_\Phi$ (here, we replaced $\bar{\Phi}$ by Φ for notations simplicity),

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

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

2.2.2 Maximum entropy models

With the KL divergence as a notion of entropy for point processes, we can now define the macro-canonical and micro-canonical models. These models are distributions of maximum entropy under different types of constraints. When considering these models as approximations of 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 $d < \infty$ (one can think of, for instance, estimators of the k nearest neighbours distribution functions, $D_k(r)$, such as in Tscheschel and Stoyan (2006)).

Macro-canonical model The macro-canonical model is defined as the distribution $\mathscr{L}$ of a point process Ξ on $\mathbb{M}^s$ that *minimizes* the KL divergence $\mathrm{KL}(\mathscr{L}, \mathscr{L}_0)$ under *expectation constraints*: $\mathbb{E}(K(\Xi)) = a$, for some vector of constraints, e.g. $a = \mathbb{E}(K(\bar{\Phi}))$ or $a = K(\bar{\phi})$. Under some technical assumptions (in particular, having density with respect to the reference Poisson distribution), the solution of the macro-canonical model is given by the Gibbs point process (Dereudre 2019, Section 1.3). Sampling from the macro-canonical model is usually computationally very expensive (Bruna and Mallat 2019). Therefore, we shall focus on the micro-canonical model, defined in the following.

Micro-canonical model The micro-canonical model is defined by replacing the expectation constraints $\mathbb{E}(K(\Xi)) = a$ with pathwise constraints. Let $\bar{\phi} \in \mathbb{M}^s$ be our observation sample, of unknown distribution. For all $\mu \in \mathbb{M}^s$, we define the energy of μ as:

$$E_{\bar{\phi}}(\mu) := \frac{1}{2}|K(\mu) - K(\bar{\phi})|^2. \tag{4}$$

The micro-canonical set of level ε, for some $\varepsilon > 0$, is defined as

$$\Omega_\varepsilon := \{\mu \in \mathbb{M}^s : E_{\bar{\phi}}(\mu) \leq \varepsilon\}. \tag{5}$$

The micro-canonical model is defined as the distribution $\mathscr{L}$ that minimizes the KL divergence with respect to the reference distribution $\mathscr{L}_0$ under *pathwise constraints* requiring $\mathscr{L}$ to be supported on Ω_ε:

$$\arg\min_{\mathscr{L}} \quad \mathrm{KL}(\mathscr{L}, \mathscr{L}_0) \tag{6}$$

$$\text{given} \quad \int_{\mathbb{M}^s} \mathbb{1}(\mu \in \Omega_\varepsilon)\mathscr{L}(d\mu) = 1, \tag{7}$$

where $\mathbb{1}(\cdot)$ is the indicator function. If $\mathscr{L}_0(\Omega_\varepsilon) > 0$, the solution to this problem (6) (7), is the measure $\mathscr{L}$ having a uniform density on $\mathbb{M}_s$ given by $\frac{d\mathscr{L}}{d\mathscr{L}_0}(\mu) = \frac{1}{\mathscr{L}_0(\Omega_\varepsilon)}\mathbb{1}(\mu \in \Omega_\varepsilon)$, $\mathscr{L}_0 - a.s.$

In order to consider the micro-canonical model as a good approximation of the observation distribution, one usually aims at finding K satisfying the following properties:

(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 moments $\mathbb{E}(K(\bar{\Phi}))$, should *characterize the unknown distribution* as completely as possible. It requires that K has a strong (distributional) discriminate power.

A natural framework allowing one to address (P1) and (P2) is by defining the descriptors $K = (K_1, \ldots, K_d)$ as a vector of empirical averages

$$K_i(\mu) = \frac{1}{|W_s|}\int_{W_s} f_i(\bar{S}_x\mu)\,dx \qquad \mu \in \mathbb{M}^s, \tag{8}$$

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

These properties are needed in order to have a model that reproduces typical geometric structures in Φ, and generates diverse samples.

In this paper, we shall consider that the number of points of our model in W_s is fixed. In such a case, it is customary to take the homogeneous Poisson point process distribution conditioned on having 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 $\mathscr{L}_0^n$.

Sampling from the uniform density on Ω_ε efficiently remains an open problem. In the literature on stochastic process modelling, most sampling algorithms rely on the following method: one first samples from an initial, high-entropy measure, and iteratively minimize the energy (cf. (4)) of this sample until it reaches Ω_ε. By choosing a high entropy initial measure, one hopes that the resulting model also has a high entropy. Recall that the micro-canonical model has the highest entropy supported on Ω_ε. Contrary to the classical methods in the point process literature (Tscheschel and Stoyan 2006; Koňasová and Dvořák 2021), which relies on random search, popular methods in image modelling use gradient descent to perform fast sampling in the micro-canonical set. However, optimizing the values of the image pixels does not guarantee that the resulting sample is an atomic measure. For these reasons, in what follows we propose a model based on the transport of a Poisson point process via a gradient descent algorithm.

3 Particle gradient descent model

In Section 3.1, we introduce the particle gradient descent model, that uses gradient descent on the positions of the particles of the sample. This model consists in using the gradient of a prescribed energy to move the points of an initial random configuration, until we obtain a pattern similar (in an informal sense) to the observation. We then present in Section 3.2 a theorem stating that this model preserves some basic invariances of the original distribution. This result, allowing us to gain some understanding about the entropy of our model, extends the results of Bruna and Mallat (2019). In Appendix C, we present some ideas about how to relax the hypotheses made in this paper, in order to build a model better suited for real world data.

3.1 Particle gradient descent model

As in Section 2.2, let $\bar{\phi} \in \mathbb{M}^s$ be our observation sample of unknown distribution, and $K : \mathbb{M}^s \longrightarrow \mathbb{C}^d$, for some $d < \infty$ a mapping defining our descriptors. We note the resulting energy $E_{\bar{\phi}}$ (cf. (4)).

Let $\bar{\phi}_0$ sample from an initial distribution that we choose as $\mathscr{L}_0^{\bar{\phi}(W_s)}$ (i.e. the same number of particles as $\bar{\phi}$, drawn uniformly and i.i.d.). We minimize the energy of $\bar{\phi}_0$ through its gradient with respect to the particles positions. More precisely, we define the mapping

$$\begin{array}{rccc} F: & \mathbb{M}^s & \longrightarrow & \mathbb{M}^s \\ & \mu = \sum_i \delta_{x_i} & \longmapsto & \displaystyle\sum_i \delta_{x_i - \gamma\nabla_{x_i}E_{\bar{\phi}}(\mu)} \end{array} \tag{9}$$

for some gradient step $\gamma > 0$. The measure $F(\mu)$ can be seen as the push-forward $F_{\mu\#}\mu$ of the measure μ by the mapping $F_\mu(x) := x - \gamma\nabla_x E_{\bar{\phi}}(\mu)$ (see e.g. Molchanov and Zuyev (2002) for more details about steepest descent methods on spaces of measures). Note that the function F_μ depends on the measure μ which is pushed forward. For any initial point measure $\bar{\phi}_0 \in \mathbb{M}^s$ we define the successive point measures:

$$\bar{\phi}_n := F_{\bar{\phi}_{n-1}\#}\bar{\phi}_{n-1}, \qquad n \geq 1. \tag{10}$$

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 $\mathscr{F}_\#$: For a probability law $\mathscr{L}$ on $\mathbb{M}^s$ $\mathscr{F}_\#\mathscr{L}(\Gamma) := \mathscr{L}(\{\mu \in \mathbb{M}^s : F_{\mu\#}\mu \in \Gamma\})$, for any $\Gamma \in \mathscr{M}^s$. Then, for an initial probability law $\mathscr{L}_{\bar{\phi}_0}$ on $\mathbb{M}^s$ we define the successive probability laws

$$\mathscr{L}_{\bar{\Phi}_n} := \mathscr{F}_\#\mathscr{L}_{\bar{\Phi}_{n-1}}, \qquad n \geq 1. \tag{11}$$

Note that $\mathscr{L}_{\bar{\Phi}_n}$ is the distribution of the point process $\bar{\Phi}_n$ obtained by n iterations of (10) starting from $\bar{\Phi}_0$ having law $\mathscr{L}_{\bar{\Phi}_0} = \mathscr{L}_0^{\bar{\Phi}(W_s)}$. Our model is defined by setting a fixed number of iterations as a stopping rule.

Observe that our model takes inspiration from the micro-canonical model, however there is no guarantee that the optimization reaches Ω_ε (defined in (5)), for any $\varepsilon > 0$. By setting a fixed number of iterations and not rejecting any configuration, we make the implicit assumption that our model reaches a low energy level. In practice, one can use classical line-search methods in the optimization to adjust the γ in (9), so as to ensure that the energy decreases as n grows.

3.2 Leveraging invariances

One can leverage some a priori known invariance properties of Φ (for instance stationarity or isotropy), by building a model that satisfies the same invariance properties as Φ. By using the descriptor K with the same invariance, the particle gradient model respects these invariance properties. In particular, we obtain a stationary point process model when K is defined by the empirical averaging (8).

This requires some explanation, since invariance properties of the distribution of Φ do not, in general, imply any natural invariance of its restriction $\bar{\Phi}$ 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. Indeed, in what follows we shall formulate a result saying that, when K and the distribution of $\bar{\Phi}_0$ 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 $\{-1, 0, 1\}$ and $x_0 \in W_s$. Note that the matrix A is restricted in integer entries for T to be a well defined invertible operator. It encapsulates translations, flips, and orthogonal rotations.

We say that:

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

Theorem 1 *Let T be a rigid circular transformation. Let $\bar{\Phi}_0$ be a point process on W_s such that its distribution $\mathscr{L}_{\bar{\Phi}_0}$ is invariant to the action of T and let K be a descriptor invariant to the action of T. Then, for all $n \in \mathbb{N}$, $\mathscr{L}_{\bar{\Phi}_n}$ defined as the push-forward of $\mathscr{L}_{\bar{\Phi}_0}$ by* (11) *is invariant to the action of T.*

A proof of the above result is given in Appendix A. This property can guarantee distributional symmetries in our model with respect to the original distribution, not stated in the classical approach of Tscheschel and Stoyan (2006). The result itself is inspired from Bruna and Mallat (2019), where the preservation of invariance is proven for the gradient descent model in the pixel domain. 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 (8) with periodic boundary condition. This boundary condition means the use of the shift operator $\bar{S}_x$ in (8), which can be interpreted as a torus correction on W_s.

A drawback is that such a boundary condition introduces a statistical bias to the spatial average (8) as an estimator of $\mathbb{E}[K(\bar{\Phi})]$ in the case of a non periodic Φ over W_s. One can expect, however, that when the window size is large enough and spatial correlations of the patterns are not too large, this border effect becomes negligible.

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 the pair

correlation function, such as Ripley's K-function, Besag's L-function, or the radial distribution function (Chiu et al. 2013, 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 Chiu et al. (2013, Section 2.3.4 and 4.1.7)). In Tscheschel and Stoyan (2006), the authors advocate the use of the k-nearest neighbors distribution function, with a k significantly greater than 1. In addition to being non-differentiable, these moments suffer from another drawback. If one wants to capture geometric structures formed by the particles, up to a fixed scale, the number of moments (i.e. the k nearest neighbours) will grow linearly with the number of particles forming such structures. This can become a problem if the intensity of the process is large, both computationally, and from a statistical point of view, as the variance of the moments may become large when estimated from a single observation.

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 the wavelet transform of a random configuration, are adapted from Zhang and Mallat (2021). They have shown high quality results in modelling geometric structures in texture images and turbulent flows.

We begin, in Section 4.1, by presenting wavelet transform for counting measures, and their so called phase harmonics, which are derived from complex wavelet coefficients by applying a multiplication 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 counting measures, and detail the choice of the descriptors that we use for numerical experiments.

4.1 Wavelet transforms and their phase harmonics

Informally, a wavelet $\psi : \mathbb{R}^2 \mapsto \mathbb{C}$ is a function that is localized both in the space and the frequency domains. Convoluted with an input signal, they allow one to capture its local geometric structure, at a given scale (see e.g. Mallat (2001, Section 7)). To capture information at different scales in the signal, we build a family of wavelets by rotations and dilations of the wavelet ψ. They constitute the foundation of 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. Oriented wavelets have already been considered, e.g. to analyze anisotropy properties of planar point processes (see e.g. Rajala et al. (2018)). We shall also use oriented wavelets which allow to capture edge-like geometric structures in the observation. Specifically, we choose bump steerable wavelets introduced in Mallat et al. (2020). They are defined by the translations, dilations and rotations of a complex analytic function $\psi(x) \in \mathbb{C}$ with $\int \psi(x)\,dx = 0$ and $\int |\psi(x)|\,dx < \infty$.

In what follows we first define the wavelet transform for the counting measures in $\mathbb{M}^s$ which are constructed from the function ψ. Let us denote the Fourier transform of ψ for $\omega \in \mathbb{R}^2$ by $\widehat{\psi}(\omega) = \int \psi(x) e^{-i\langle \omega, x\rangle} dx$. By construction, 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 spatial 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 $0 \le \ell < L$, 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

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

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 has negligible amplitude for $|x| > 2^j C$.

For a counting measure $\mu \in \mathbb{M}^s$, we typically consider only the wavelets having spatial support [1] contained in W_s by limiting the scale $j < J$ such that $2^J C \le 2s$. 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 $0 \le j < J$, $0 \le \ell < L$, 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 $\mu \in \mathbb{M}^s$ observed in the finite window W_s, is a family of functions obtained by the convolution of μ with periodic wavelets ψ^s_λ,

$$\mu \star \psi_\lambda(x) = \int_{W_s} \psi^s_\lambda(x-y)\mu(dy), \quad \lambda \in \Lambda. \tag{12}$$

They are defined with *periodic edge connection*, i.e. at $x = (x_1, x_2) \in W_s$, $\psi^s_\lambda(x_1, x_2) := \sum_{n_1, n_2 \in \mathbb{Z}} \psi_\lambda(x_1 + 2sn_1, x_2 + 2sn_2)$. The integral (12) can be interpreted as a shot-noise, which is thus well-defined because $\int_{\mathbb{R}^2} |\psi_\lambda(x)| dx < \infty$. We denote the wavelet coefficients of $\mu \in \mathbb{M}^s$ by $\{\mu \star \psi_\lambda(x)\}_{\lambda \in \Lambda, x \in W_s}$.

Remark: As the wavelet transform is a linear transformation of a counting measure, it is known that the covariance between $\bar{\Phi} \star \psi_\lambda(x)$ and $\bar{\Phi} \star \psi_{\lambda'}(x')$ depends only on the

[1] More precisely, where the wavelet norm is non negligible

mean intensity and second-order correlations of a stationary point process Φ (Brémaud 2002, Eq. (5.27)), which only gives partial information on the process distribution.

4.1.2 Wavelet phase harmonics

To capture random geometric structures of different scales occurring simultaneously (i.e. at nearby x and x') in a given process, one can compute the covariance of the wavelet coefficients at different scales. However, due to the frequency localization property of the wavelets, such covariance can be close to 0, even though the wavelet coefficients are not independent. To capture such dependencies, one can use a non linear operator on the transforms, to superimpose their frequency support. This section, along with the following, details this non linear operator, and the resulting covariance moments.

Phase harmonics (Mallat et al. 2020) 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 $x \in W_s^2, \lambda \in \Lambda$, $k \in \mathbb{Z}$, let's denote the *wavelet phase harmonics* of $\phi \in \mathbb{M}^s$ 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. Note that the wavelet phase harmonics at $k = 1$ are exactly the wavelet coefficients .

As illustrated in Zhang and Mallat (2021), 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 nonlinear frequency transposition property is crucial to capture dependencies of the wavelet coefficients across scales and angles, as we shall detail next.

4.2 Wavelet phase harmonic covariance descriptors

A classical way to capture dependencies between wavelet coefficients is to compute their higher order moments. However, as the order grows, 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 do 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 correlations, while still capturing information beyond second-order correlations (to satisfy (P2)).

The wavelet phase harmonic covariance of $\bar{\Phi}$ is defined by

$$\mathrm{Cov}([\bar{\Phi} \star \psi_\lambda(x)]^k, [\bar{\Phi} \star \psi_{\lambda'}(x')]^{k'}), \tag{13}$$

for pairs of $(x, x') \in W_s \times W_s$, $(\lambda, \lambda') \in \Lambda^2$, 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 Mallat et al. (2020); Zhang and Mallat (2021), for a stationary process Φ, the overlap between the frequency support of $[\bar{\Phi} \star \psi_\lambda]^k$ and that of $[\bar{\Phi} \star \psi_{\lambda'}]^{k'}$ 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 second order correlations). We shall also follow this empirical rule to select a covariance set Γ_H (specified in detail in Section 4.3) to describe point processes.

Let $v_{\lambda,k} = \mathbb{E}([\bar{\Phi} \star \psi_\lambda(x)]^k)$. We define the descriptors $K(\mu)$ using (8), as empirical estimators of moments. Additionally, let us denote $\mu_{\lambda,k}(x) := [\mu \star \psi_\lambda(x)]^k - v_{\lambda,k}$. Taking the spatial average (8) gives the descriptor of the form:

$$K(\mu) = \left(\frac{1}{|W_s|} \int_{W_s} \mu_{\lambda,k}(x)\, \mu_{\lambda',k'}(x - \tau')^* dx \right)_{(\lambda,k,\lambda',k',\tau') \in \Gamma_H}. \tag{14}$$

As $\bar{\Phi}$ is circular-stationary, (13) depends only on $x - x'$, it suffices to use the vectors τ' to measure the differences between x and x'. Note also that $K(\mu)$ 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 (14) by $\bar{v}_{\lambda,k} = \frac{1}{|W_s|} \int_{W_s} [\bar{\phi} \star \psi_\lambda(x)]^k dx$ 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 (13). This is a good approximation of $K(\mu)$ as the estimation variance of the covariance moments is typically much larger than that of the first-order moments.

4.3 Choice of the covariance set Γ_H in (14)

Rather than detailing the full list of elements in Γ_H, we provide an intuitive way to choose the set Γ_H. For the full list, see

Brochard et al. (2020, Section 4). Overall, the total number of elements in Γ_H is in the order of $O(L^2J^2)$. Note that the smallest structures that the descriptors can capture depend on the spatial support C of the wavelet ψ. Information about structures smaller than C can be added in a post-processing step will be explained in Section 5.

- Choice of J: The covariance set Γ_H depends on the parameter J, which is the maximal scale of the wavelet transform. A suitable choice for this parameter J would be one allowing for a good trade-off between satisfying the sufficiency of K, while maintaining the concentration property (cf. properties (P1) and (P2) from Section 2.2).
- The parameter τ' is chosen so that each wavelet is translated in a particular direction in order to capture correlations along nearby edges in the observation.
- Choice of $(\lambda, \lambda', k, k')$: These parameters are chosen in order to capture 2nd-order correlations, as well as dependencies between wavelet coefficients at different scales and orientations, both with and without phase information, based on a rule of thumb that $k\lambda \approx k'\lambda'$ due to the frequency transposition property of the wavelet phase harmonics. Figure 2 shows the impact of using phase harmonics coefficients (with $k, k' \neq 1$) compared to 2nd-order correlations (only $k = 1, k' = 1$). We see that both syntheses deviate from complete spatial randomness, but important structures, such as vortexes, are better reproduced when incorporating the non-linear coefficients.

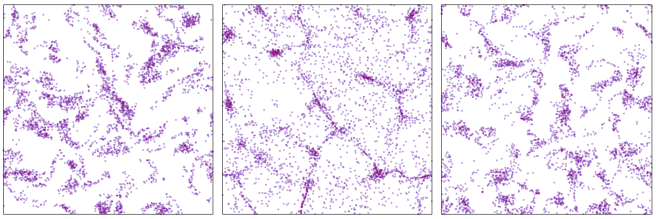

Fig. 2 Left: Observation of a turbulence Poisson process (original), middle: synthesis with wavelet covariances without phase harmonics non-linearity (i.e. only $k = k' = 1$ in Γ_H in (14)), right: synthesis with wavelet phase harmonic covariance descriptors (full Γ_H)

5 Numerical scheme for particle gradient descent

Calculating the wavelet phase harmonic covariances can be computationally demanding (due to the calculation of two integrals). In order to gain some efficiency, we can perform the computations in a discrete domain. However, the energy needs to remain differentiable with respect to the positions of the points in the pattern. We propose a method, consisting of a Gaussian smoothing of the configuration of points, to address this problem. Building on that method, we then present two technical aspects of the sampling method.

In this section, we discuss a complete numerical scheme to generate samples from the particle gradient descent model, defined with the wavelet phase harmonic descriptors presented in Section 4. It is composed of the following ideas:

- *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. At each scale, we use a quasi-Newton gradient-descent method for greater efficiency.
- Final *blurring* (optional): to add a priori information on structures whose size is smaller than C into the model samples. It helps 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 (14) 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 (12), 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 the automatic differentiation software Pytorch (Paszke et al. 2019) to perform the following operations. It allows one to compute the derivative of a modified energy w.r.t. any point x_i in μ. The following paragraph details this discretization:

We first map a given point measure μ on W_s to a continuous function μ_σ by the convolution

$$\mu_\sigma(x) := \mu \star g_\sigma(x) = \sum_{x_i \in \mu} e^{-\frac{|x-x_i|^2}{2\sigma}}, \quad x \in W_s, \tag{15}$$

with a (periodized) Gaussian function g_σ of given standard deviation σ. Then we evaluate μ_σ on the $N \times N$ regular grid inside W_s and denote the resulting matrix μ_σ^N, with entries called (values of) *pixels*. The convolution with a Gaussian function makes each entry of μ_σ^N smoothly depend on the atom positions of μ. We then compute $\bar{K}(\mu_\sigma^N)$ instead

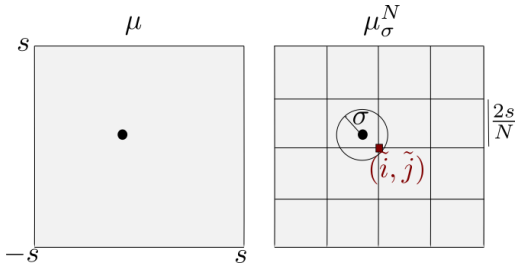

Fig. 3 The differentiable discretization of a courting measure μ into a point-image μ_σ^N on a $N \times N$ regular grid.

of $K(\mu)$, where $\bar{K}$ is this discrete analogy of the descriptor (14) (cf. Zhang and Mallat (2021)). Note that, because the value of a pixel continuously depends on the positions of the atoms, this discretization makes our descriptor only invariant to discrete translations (multiple of the pixel size $2\frac{s}{N}$), for which Theorem 1 applies. The gradient of the energy $|\bar{K}(\mu_\sigma^N) - \bar{K}(\bar{\phi}_\sigma^N)|^2$ with respect to each atom position of μ can therefore be computed using automatic differentiation (with the Pytorch software). Indeed, we know that $\bar{K}$ is differentiable w.r.t. each entry of μ_σ^N, as a combination of linear and non-linear operators. Moreover, for any $i, j \in \{1, N\}^2$, noting $\tilde{i} = -s + 2si/N$, $\tilde{j} = -s + 2sj/N$, (15) gives us that

$$\mu_\sigma^N(i,j) = \mu \star g_\sigma(\tilde{i},\tilde{j}) = \sum_{x_i \in \mu} e^{-\frac{|(\tilde{i},\tilde{j}) - x_i|^2}{2\sigma}}, \tag{16}$$

which is differentiable w.r.t. any x_i in μ. This discretization step is illustrated in the Figure 3.

In signal processing, the Gaussian function acts as a low-pass filter. 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.

5.1.2 Wavelet discretization and choice of scales

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 N, we set $C = \frac{2s}{N}$. In this way, the spatial support of ψ has a radius C of one pixel of the image. As a consequence, this smallest-scale wavelet ψ can also be discretized (without significant aliasing) in order to compute the discretized descriptor $\bar{K}$.

The choice of the largest scale 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 (large values of $|\lambda|, |\lambda'|$ in (14)), due to the fact that the point-images are composed of local spikes when σ is small. This implies that these high frequency statistics have an important impact on the gradient of $\bar{K}$, 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 optimization issue can be overcome by matching the descriptors from low frequency to high frequency in a sequential order, through an appropriate modulation of the parameter $\sigma \in \{\sigma_0, \sigma_1, ..., \sigma_{J-1}\}$ 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 (15) 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 σ, 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 (small values of $|\lambda|, |\lambda'|$ in (14)) are matched with the ones of our observed sample. Thus, we propose a *multiscale* gradient descent procedure that 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 until $\sigma = \sigma_{J-1}$). We choose $\sigma_j := \frac{s}{N} 2^{J-j-2}$. Note that σ_{J-1} is equal to $\sigma_{\min}$. For numerical efficiency, we perform the gradient descent procedure using the L-BFGS optimization algorithm (Liu and Nocedal 1989).

5.3 Final blurring

We observed that the contrast between the continuous nature of our objects and the discrete approximation described in Section 5.1 creates undesired artificial structures at frequencies higher than the image resolution: when the number of particles in a configuration is large with respect to the number of pixels in the image, or if the configuration exhibits strong clustering behaviour, several pixels may contain more than one particle. In such cases, our algorithm produces samples having an artificial clustering structure *inside* each of these

pixels (see Brochard et al. (2020, Section 5) for an illustration of this phenomenon).

To remove this artificial clustering, 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.

6 Numerical experiments

In this section we present 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 evaluate these models by comparing samples from the original distributions to samples from our models, visually as well as by estimating their power spectrum, which we define in Appendix B. The power spectrum gives information equivalent to the second order correlation function of the process (Brémaud 2002), which captures clustering or repulsive behaviour between atoms of a realization. Such information cannot always be detected visually. In order to further quantify how well our model captures visual geometric structures, and to gain some insight into the ability of our model 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.

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 distributions

For our experiments, we choose point process distributions that show complex geometric structures, for which we can visually recognize geometric structures. 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, see e.g. Skare et al. (2007), and the Boolean model with circular grains of fixed radius, considered in Chiu et al. (2013, 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.

Then, we take interest in distributions having *turbulent intensity* (derived from the simulations of a decaying isotropic turbulent vorticity field driven by 2d Navier-Stokes equations, see e.g. Schneider et al. (2006)). Such fields 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. From the turbulent intensity, we sample three different processes, exhibiting distinct microscopic structures (repulsive, independent or clustering): a Matern *cluster* process, a Poisson point process and a Matern II *hard-core* process, see Chiu et al. (2013, 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.

The number of points in the Cox Voronoi, Small circles, Big circles, and the Turbulent Hardcore, Poisson, ad Cluster processes are around, respectively, 1 900, 2 500, 2 000, 1 700, 3 800 and 13 000. Note that, for comparison, point patterns considered in Tscheschel and Stoyan (2006) have around 400 points.

6.1.2 Choice of model parameters

Image resolution 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. For simplicity, we fix $s = 1/2$ for all the examples that we shall consider. The ultimate Gaussian variance (precision) of this mapping is thus $\sigma_{\min} = \frac{1}{2N}$. 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 one example where a higher resolution, $N = 256$, is used to capture most of the high frequency information.

Number of iterations The number of iterations of the L-BFGS optimization is chosen to be 100 for each scale σ_j (a total of 400 iterations for $N = 128$, and 500 iterations for $N = 256$).

Other parameters and computation time Empirical evidence in Section 6.2 shows that the multi-scale optimization procedure in Section 5.2 allows one to reconstruct (modulo translation) the observed sample when using $\bar{K}$ defined with $J = \log_2(N) - 2$, 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 also apply this multiscale optimization method to our model defined with $J = \log_2(N) - 3$. The number of angles in the steerable wavelets is $L = 8$.

An overview of the main parameters of our model is given in Appendix D. The average computation time on 4 GPU (Nvidia Tesla P100) for a sample for a turbulent process having roughly 13 000 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 (property (P2)) is visual evaluation, which is widely used in image analysis but subjective. We then compare the power spectra (cf. Appendix B for the definition) of our models and the original distributions. To estimate the power spectra, we generate (for each original distribution) 10 i.i.d. samples from the same model (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 original distribution. All these samples will also serve in Section 6.3 to compare their geometric similarities.

Figure 4 shows a study of our three Cox distributions. The first line presents samples from the original distributions. The second line presents samples from the model using our descriptor with $J = \log(N) - 2$. In this setup, the concentration property is not satisfied (see property (P1) in Section 2.2.2), and the result is the memorization of the observation sample $\bar{\phi}$. Indeed, this line shows quite faithful reconstructions of the original samples subjected to a periodic translation, up to some precision error due to a finite image resolution N. This is however not a good model because it essentially only contains the observation $\bar{\phi}$. It suggests that we need to improve the concentration property (P1) of the descriptors in order to enlarge the ensemble Ω_ε. Note that, in the work of Tscheschel and Stoyan (2006), the authors use the term 'reconstruction' to refer to random sampling method, which we call in this paper 'synthesis'.

In order to improve (P1), we shall reduce the parameter J in the wavelet transform. The third line of Figure 4 shows realizations sampled using $J = \log(N) - 3$ for different original distributions. Our analysis in Section 5.1.2 suggests that this range of J can model structures whose spatial size is at most 1/8 of the window W_s. Observe that most polygons and circles are well reproduced in the synthesis of Voronoi and Small circles. The Big Circles are harder to model since the size of each circle is slightly larger than 1/8 of W_s.

In the last line of Figure 4, we present the power spectra for $k \in \mathbb{N} \cap [1, 128[$ (cf. Appendix B) from the original distributions and as well as from our model. Larger errors can be observed at k near zero (say k=1,2 and 3). This is because only the average spectral information is captured (and matched) using the low-pass filter ψ_0 in the wavelet transform, which is included in the descriptor $K(\mu)$ (c.f. (14)). Moreover, the variance of the empirical information at small k can create extra error since it can be far away from its expectation. Similarly, because the wavelet convolutions average the spectral information over different frequency bands when using a reduced number of τ' in $K(\mu)$, our descriptor does not capture fast oscillations in the power spectra in the range of $k \leq N/2 = 64$ (see Zhang and Mallat (2021) for more details about how to capture these oscillations). This is observed in the cases of Small and Big Circles Cox processes. See Brochard et al. (2020, Appendix C) for a theoretical formula of the power spectrum in the case of Small and Big circles. When $k > N/2 = 64$, the descriptor $\bar{K}$ (cf. 5.1) does not contain accurate spectral information due to a finite image resolution N. In this regime, we observe a smooth decay of the (log) power spectrum towards 0. We observed that if we apply the final blurring (cf. Section 5.3), then the error of the model spectrum becomes larger. Therefore, for these three processes, no final blurring has been applied.

All models discussed up to now are Cox processes, with Poisson (hence independent) points sitting on some random macroscopic structures. Figure 5 presents our analysis of three turbulent point processes having different microscopic structures: a hardcore, a non-correlated (Poisson) and a clustering one. We see that our generated samples capture to some extent this microscopic structure. For the clustering model, the presented synthesis is done with $N = 256$. We see that our model (using $J = \log(N) - 3$) can generate samples with similar macroscopic and microscopic structure. The power spectrum at small k has larger errors, as we have observed in the Cox models. However, since the power spectra are mostly smooth in these Turbulent processes, we observe a relatively small spectrum error over a wide range of k. This is also due to the use of the final blurring which helps to remove some artificial spectrum errors for $k > N/2$. For the clustering model, we also compare the power spectrum of two models with different resolutions: $N = 128$ and $N = 256$. We see that setting a higher resolution reduces significantly the error, allowing to match the spectrum up to $k \simeq 80$. we still observe some small error when $k \geq 80$, probably because

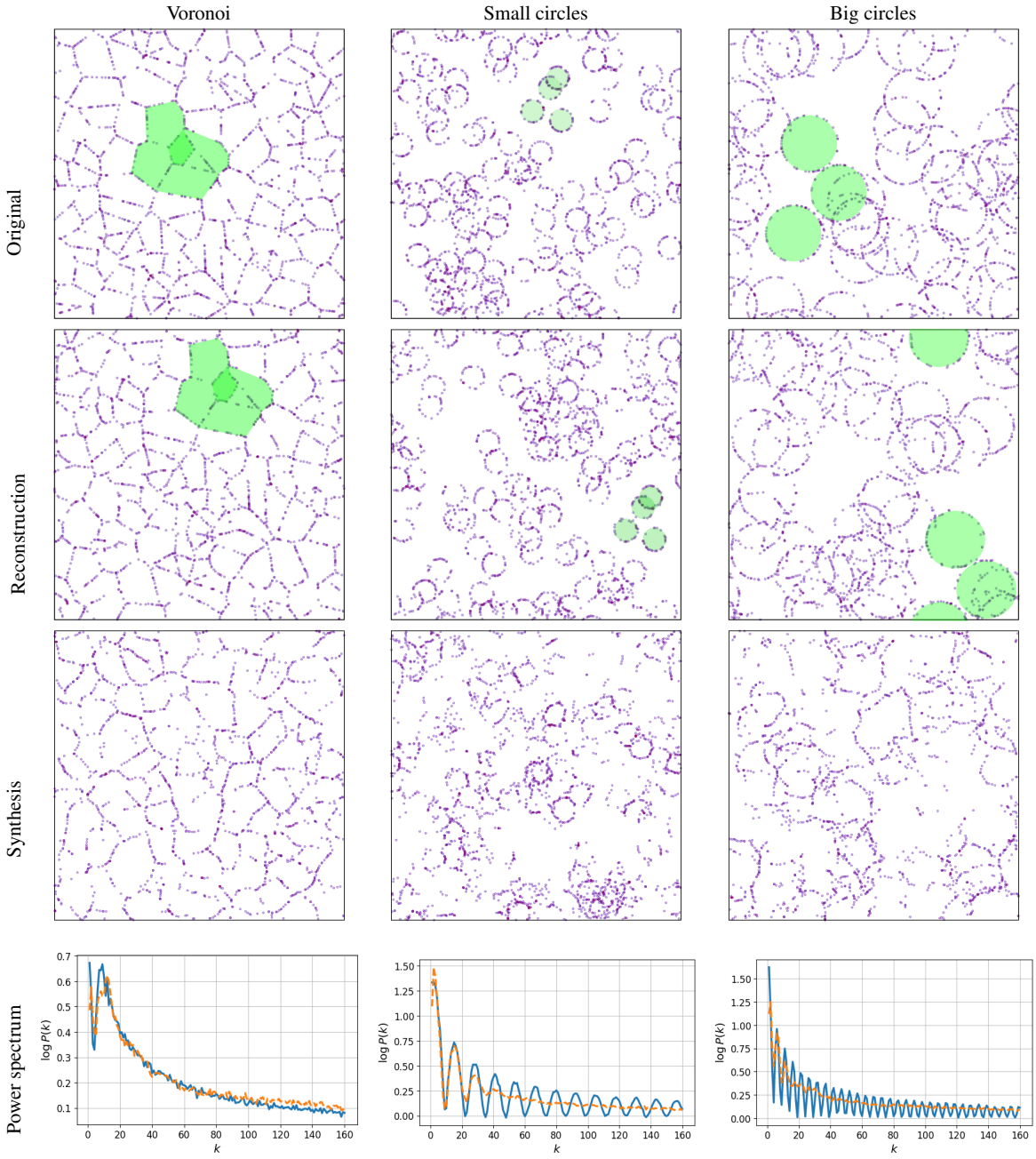

Fig. 4 Three Cox models; original sample, reconstruction, and synthesis. For the power spectrum plots, the full lines (in blue) correspond to the original distributions, and the dashed lines (in orange) correspond to the models. The y-axis is presented in log scale.

of the final blurring (cf. Section 5.3), which may also impact the high frequencies that we optimize. Overall, both the visual and the spectral analysis suggest that our model can generate well various Turbulent points processes.

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, 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

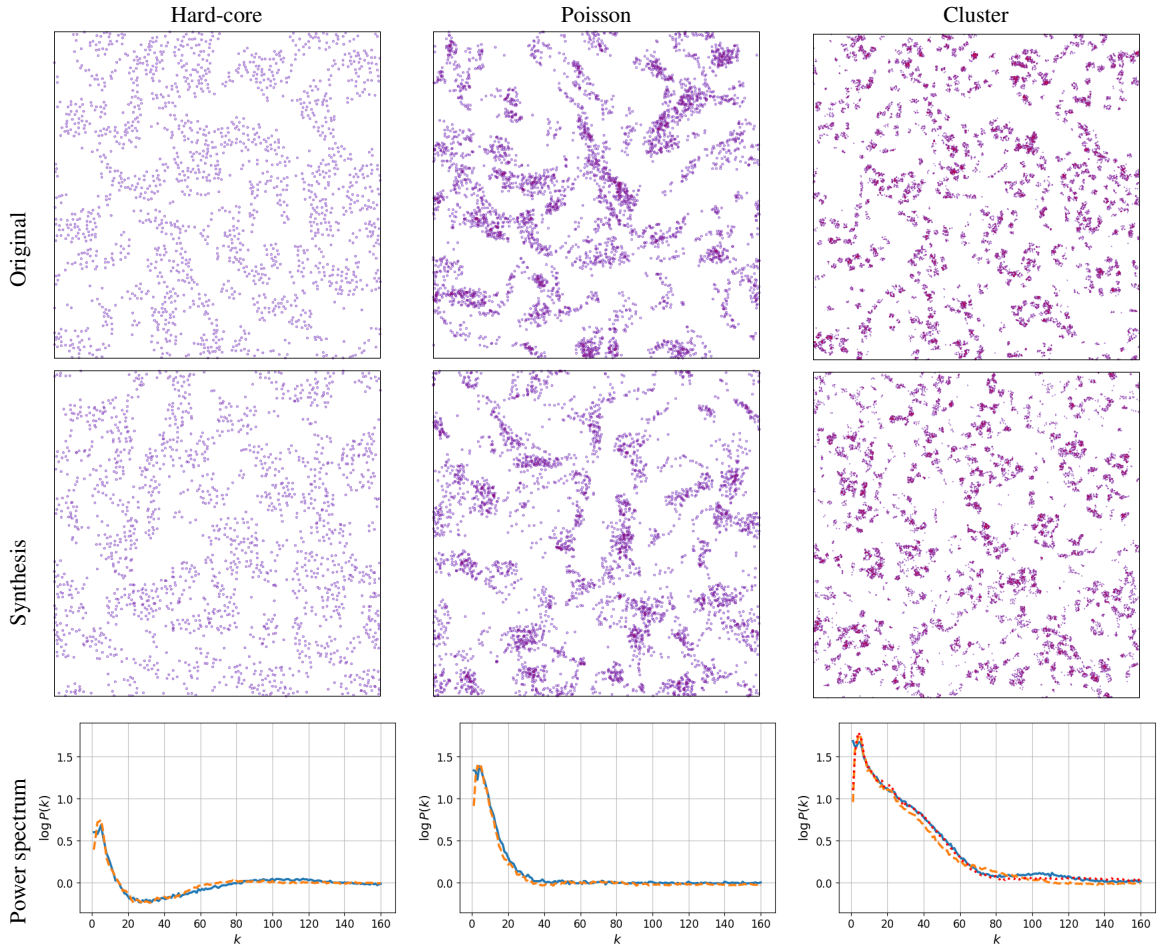

Fig. 5 Synthesis of turbulence processes with various microscopic structures (Hardcore, Poisson, and Cluster), and their power spectrum plots. Full lines (in blue) correspond to the original processes, dashed lines (in orange) correspond to the model. For the Cluster distribution, the dotted line (in red) corresponds to our model defined with the resolution $N = 256$.

evaluate in a simple way the ability of our model to produce diverse samples.

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 Boissonnat et al. (2018, 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 Brochard et al. (2020, Appendix C) for more details about this link).

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.[2] 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 small deformations of ϕ. It is hence interesting to use this alternative tool to evaluate our generative model.

[2] In our case we use the periodic metric.

Topological data analysis Our approach in this matter is inspired by Chazal and Michel (2017), and we refer the reader to this paper for a more detailed description. We use the 'holes' birth-death process, as it appears more relevant to capture information in the Cox distributions, such as the polygons and the circles.

In order to compare the distributions 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 between two point clouds on the plane (we found that the bottleneck distance, also suggested in Chazal and Michel (2017), is not sufficiently discriminating for our point patterns). 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 visualize the representation of all samples.

TDA of our experiments In the plots on the first line of Figure 6, we study separately the Cox Voronoi and Big circles processes, and the turbulent hardcore process. In each plot, we observe 20 *dots* (having different shapes), each representing one 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 and 10 representing realizations from the generative model. For each plot, the sample additionally marked with a black dot represents the observation used in our model to produce the 10 syntheses.

We see in the first two Cox examples a clear separation between the original process and the model, implying a lack of sufficiency (P2) in our model. This is probably because, in order to satisfy (P1) and produce diverse samples, we have chosen to reduce J, and therefore lose some information about large scale structures of the process. On the other hand, we observe that this separation is smaller for the Turbulence hard-core case, where there is a better balance between (P1) and (P2). The error in the Cox models is probably due to the difficulty to reproduce highly constrained structures (perfect circles or convex polygons). These observations agree with our visual evaluation of the syntheses: the Voronoi and (more particularly so) the Big circles models are easily discriminated from their original distributions (their highly constrained structures are not perfectly reproduced in the syntheses), but this discrimination is harder for the Turbulent hardcore case. Moreover, these figures indicate, by the spread of the dots representing the syntheses, that our model reproduces, to some extent, the diversity in the samples of the original distributions (suggesting a certain entropy in our model). To further reduce the distances between the model samples and the original samples of a process, while maintaining a similar diversity (and hence a similar entropy between the model distribution and the original distribution) remains an interesting problem for future works.

The two plots in the second line present the TDA of the three Cox distributions together, and the three turbulent distributions together. We observe that, for the Cox distributions, the Small circles model is about as close to the original Small circles distribution as it is to the original Big circles distribution. Nevertheless, the original distribution of the Big circles and the Small circles are well separated, suggesting that there is a topological distinction that is not well respected in the small circles. However, this kind of error is hard to perceive visually. On the other hand, the Voronoi case is well separated from the other two, suggesting that the model is better than the ones of the circles distributions. For the turbulence case, we observe that the three distributions (both original and model) are well separated, and each respective model is closer to its original distribution than to the other distributions. This agrees with our earlier visual and spectral analysis, suggesting that our model is able to capture complex geometric structures formed by a large amount of points.

It remains an open question to quantify the influence of the range parameter on the distances between patterns. We chose to include all radii ($0 \leq r \leq \frac{1}{2}$), as we want to include information pertaining to individual point patterns, in order to measure the diversity in the distribution. In order to get an idea of the influence of this range parameter, it is possible to look at the Euler-Poincaré characteristic (see e.g. Illian et al. (2008, Chapter 4)), which is closely linked to TDA (See Section 7 for further discussions). Other related methods to compute distances between point patterns, such as in Müller et al. (2020), could constitute an interesting line of research for other evaluation methods.

Remark: We used the R packages `TDAstats` (Wadhwa et al. 2018) to calculate the PDs of our point patterns and `TDA` (Fasy et al. 2014) to calculate their Wasserstein distances. Due to memory constraints, for second line, the analysis was done using a random thinning to reduce the number of points of each sample to 2 000, which could artificially impact the results. The experiments were repeated several times and the variability in the random thinning did not impact our conclusions.

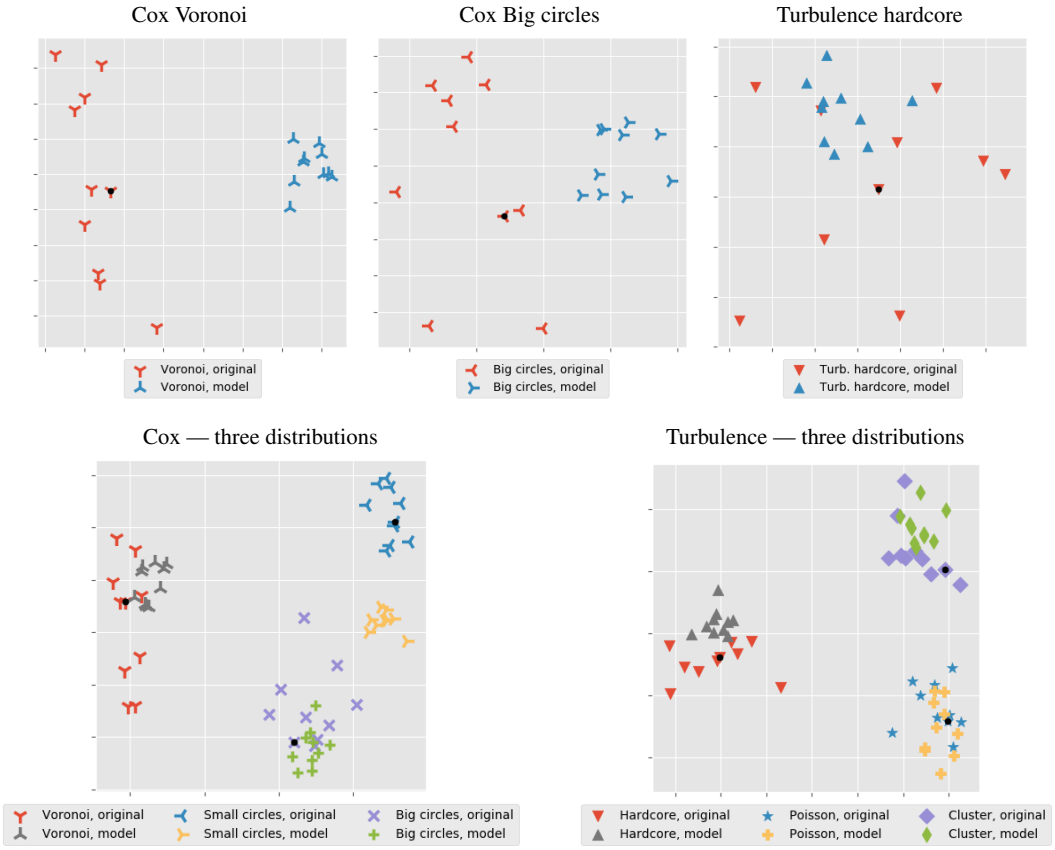

Fig. 6 TDA of the distributions presented in this paper, and their respective models.

7 Comparison between our method and Tscheschel and Stoyan (2006)

In this section, in order to illustrate the advantages of the method presented in this paper, we present a brief comparison between the method presented in Tscheschel and Stoyan (2006) and ours.

7.1 Differences between the two methods

Both methods are based on the following idea: to produce similar but different point patterns to a given observation, one first defines what should be 'similar' between the observation and the synthesis, by choosing a set of statistical constraints, computed on the observation. Then, starting from an initial random configuration of points, one iteratively modifies this configuration in order to match the set of prescribed statistics (by minimizing an energy E, related to the square difference between the statistics of the original and the synthesised point patterns). If the set of statistics does not describe the observation itself, but rather its underlying distribution, then the output of the optimization procedure should be a new point pattern, similar but different to the observation.

However, the two methods differ on two major points:

- First, the optimization method to match the set of statistics. In Tscheschel and Stoyan (2006), the optimization steps can be described as follows: given the point configuration being synthesized $\phi_k = \sum_{i=1}^{N} \delta_{x_{i,k}}$ at some step k, a point in the configuration is chosen uniformly at random, say $x_{j,k}$, for $j \in \{1, \cdots, N\}$. A candidate for a new point $y \in W$ is chosen uniformly at random in the observation window. Then, if the energy of $\tilde{\phi}_k := \phi_k - \delta_{x_{j,k}} + \delta_y$ is lower than the energy of ϕ_k, we define $\phi_{k+1} = \tilde{\phi}_k$. Otherwise, $\phi_{k+1} = \phi_k$. We call this optimization *random search* (RS).
  In our method, $\phi_{k+1} = \sum_{i=1}^{N} \delta_{x_{i,k} - \nabla_{x_{i,k}} E(\phi_k)}$, where $\nabla_{x_{i,k}} E(\phi_k)$ is the gradient of the energy with respect to the point $x_{i,k}$ of ϕ_k, see (9). This optimization method will be noted (GD).
- Second, the set of statistical constraints used to describe the geometry of the point patterns. In Tscheschel and Stoyan (2006), the authors use the k nearest neighbour distance distribution functions (d.f.) $D_k(r)$, for $k \in \{1, \cdots, k_{max}\}$, $k_{max} \geq 1$, and evaluated at a sequence of radii $r \in \{r_0, \cdots, r_{max}\}$, $r_{max} > 0$, (Stoyan and Stoyan 1994, p. 267). We call this statistical descriptor *nearest neighbour distances* (NND).

Our statistics are based on the covariance between phase harmonics of the wavelet phase harmonics coefficients (WPH) of the point patterns (see 4).

7.2 Preliminary discussion

Before presenting a numerical comparison between the two methods, we briefly explain what the limitations of the method in Tscheschel and Stoyan (2006) are, and why our method could overcome such limitations.

The optimization method in Tscheschel and Stoyan (2006) is based on random search. This implies that for some configuration ϕ_k at some step k, there may be a lot of failing new candidates to replace some point in ϕ_k before finding one that reduces the energy of ϕ_k. This means that there may be a lot of energy evaluations before updating the current configuration. Furthermore, each iteration, requiring one energy evaluation, only moves one point in the configuration. Conversely, at each iteration, our algorithm computes the energy as well as the gradient of the energy, and all the points are moves according to the gradient. This implies that, for some energy level $e > 0$, the gradient descent method may reach this level in less iterations.

Moreover, the statistical constraints used in Tscheschel and Stoyan (2006) are based on 3 parameters: the number of neighbours for the points in the configuration, the maximal radius at which to evaluate whether or not there is a neighbour, and the number of radii between 0 and this maximal radius. While the latter relates to the precision of the d.f.'s, k_{max} and r_{max} have to be chosen carefully, so as to describe the geometry formed by the points, up to some scale. The maximal radius r_{max} can be seen as the maximal scale up to which the constraints describe the geometric structure. This can be fixed depending on the observation, but should not be too large, in order to satisfy the ergodic averaging property. However, for a fixed sequence of radii, the parameter k_{max} can change significantly depending on the observation. Even if two configuration exhibit structures up to similar scales, the number of nearest neighbours inside some ball may differ depending on the intensity of the process. For instance, consider the Cox Circles distribution, where points are located on circles of fixed radius r_0, with the center of those circles forming a Poisson point process. If the observed pattern has around 10 points per circle, one would probably need to fix $r_{max} = 2r_0$, and $k_{max} = 10$. However, if the circles contain an average of 100 points, then one would have to increase k_{max} up to 100, even if the circles have the same size as before. This would increase significantly the number statistics to compute at every step. Conversely, our descriptors only depend on a number of scales at which we compute the wavelet coefficients, which does not depend on the intensity of the process. The only parameters to fix are the maximal scale J, and the minimal scale, set by the resolution N, which relates to the precision set by the number of radii in the k nearest neighbours case. In the above Cox Circles example, if the points have Poisson distribution on the circles, then the size of the descriptor will not change between the two setups (10 or 100 points per circles).

7.3 Numerical comparison

In this section we present a numerical comparison between the two methods. This comparison aims at illustrating the three following points:

1. For the same energy, the gradient descent optimization method reaches low energy levels in less time (i.e. less evaluations of energy value) than the optimization method used in Tscheschel and Stoyan (2006). To highlight this point, we shall consider the Cox Voronoi example, and use the WPH descriptors (c.f. (14)) to define the energy, and compare the RS and GD optimization methods.
2. The amount of information captured by the kth nearest neighbours d.f.'s depends on the intensity of the process, regardless of the scales of the structures. Considering the Turbulent Poisson example with different intensities, we shall see that, for a fixed descriptor (i.e. fixed k_{max}, r_{max}, number of r), the quality of the syntheses decreases with the intensity of the process.
3. We perform an overall comparison of the two methods, on the Cox Voronoi and Turbulence Poisson examples. Besides the visual and TDA comparison, we further provide statistical performance metrics to illustrate the better performance of our method.

7.3.1 Comparison between RS and GD (Cox Voronoi example)

For this experiment, we study the Cox Voronoi example, and define the energy from the wavelet phase harmonics covariances, presented in Section 4. Let K be our descriptor (defined in (14)), $\bar{\phi}$ our observation sample, and $E_{\bar{\phi}}(\cdot) = \frac{1}{2}|K(\cdot) - K(\bar{\phi})|^2$ the corresponding energy. We define the relative energy by

$$e(\cdot) = 2\frac{E_{\bar{\Phi}}(\cdot)}{|K(\bar{\phi})|^2} = \frac{|K(\cdot) - K(\bar{\phi})|^2}{|K(\bar{\phi})|^2}. \tag{17}$$

We ran the optimization of the energy with the random search method from Tscheschel and Stoyan (2006), and observe the relative energy of the syntheses (for 10 syntheses), after $n = 19870$ and $n = 29805$ iterations, i.e. respectively 10 and 15 iterations per point. After $n = 19870$ iterations, the algorithm reaches a relative energy of $e = 9,00.10^{-4}$ (with a std of $9,23.10^{-5}$), and after $n = 29805$ iterations, we found $e = 4,76.10^{-4}$ (std= $2,47.10^{-5}$), indicating that the

optimization has reached a low energy level. It took an average of 1h04min and 1h36min respectively. We observed the respective relative energies, and ran our gradient descent optimization algorithm (without the multi-scale procedure) until the relative energy reaches the levels from the random search method. The results and comparisons with our method are summarized in Table 1. The computations have been run on a single GPU Nvidia Tesla P100.

	Random search	Gradient descent
$e = 9,00.10^{-4}$	19870 (1h04m)	52 (0m35s)
$e = 4,76.10^{-4}$	29805 (1h36m)	69 (0m45s)

Table 1 Speed comparison between random search and gradient descent, in number of iterations (computation time in parenthesis) for the synthesis of Poisson Voronoi patterns. The time per iteration in the gradient descent method is larger, due to the possible several energy (and gradient) evaluations for the line search. However, the total amount of time is much lower.

7.3.2 Dependence of WPH and NND on the intensity of points (Turbulence Poisson example)

To illustrate our second point, for NND descriptor we fix the parameters of the kth nearest neighbours d.f.'s to $k_{max} = 16$, $r_{max} = .125$ (on a window of size 1), and discretize $D_k(r)$, $r \in (0, r_{\max}]$, regularly by 250 values of radii r. With this fixed descriptor, we perform a synthesis using RS optimization for three different observations: the Turbulence Poisson observation randomly thinned to have 500 points, the same observation thinned to 2000 points, and the raw observation, which contains 3784 points. We set the number of iterations to 400 iterations per point, (which is the same for both methods). Figure 7 shows examples of syntheses for the 3 different patterns, as well as syntheses from our method using WPH descriptor with GD multiscale optimization. We observe that the method using the RS+NND fails to reproduce the geometric structures in the example with the largest number of points.

Statistical evaluation metrics we present the estimations of two statistics. The first one is the spherical contact distribution function (SCDF), defined for a point process Ξ as $H_s(r) := 1 - \mathbb{P}(\Xi \cap B(0,r) = \emptyset)$, where $B(0,r)$ denotes the ball of radius r, centered at 0. The second one is the Euler-Poincaré characteristic, defined from the persistence diagram of a point pattern (cf. 6.3) as the number of connected components minus the number of holes, in function of the radius r. For these two statistics, each radius r, and each distribution, we estimate their value by averaging over 10 realizations. A confidence interval is computed using a bootstrap method, see e.g. Efron and Tibshirani (1994), with 9999 resamples, a confidence level of .95, and the 'BCa' method. We also compare this estimation with the normalized standard deviation of our samples, under Gaussianity assumptions (see for instance Efron and Tibshirani (1994, Chapter 5)).

The curves of Figure 8 confirm our visual evaluation: as the number of points in the pattern grows large, the error (deviation from the curve of the true distribution) becomes larger for the RS+NND model. In the example with the largest number of points, the SCDF curve of this model is significantly above the curve of the true distribution, because the observation contains clusters formed by a large number of points, which is not captured by the NND descriptor with $k_{max} = 16$. Therefore, large empty regions are not reproduced, and the probability of having a point inside some ball of given radius is too high. Similarly, the curve of the Euler characteristic (χ) of the RS+NND model deviates from the one of the true distribution when the number of points is large.

To quantify more precisely these errors, we report in Table 2 values of the two statistics for several relevant radii r.

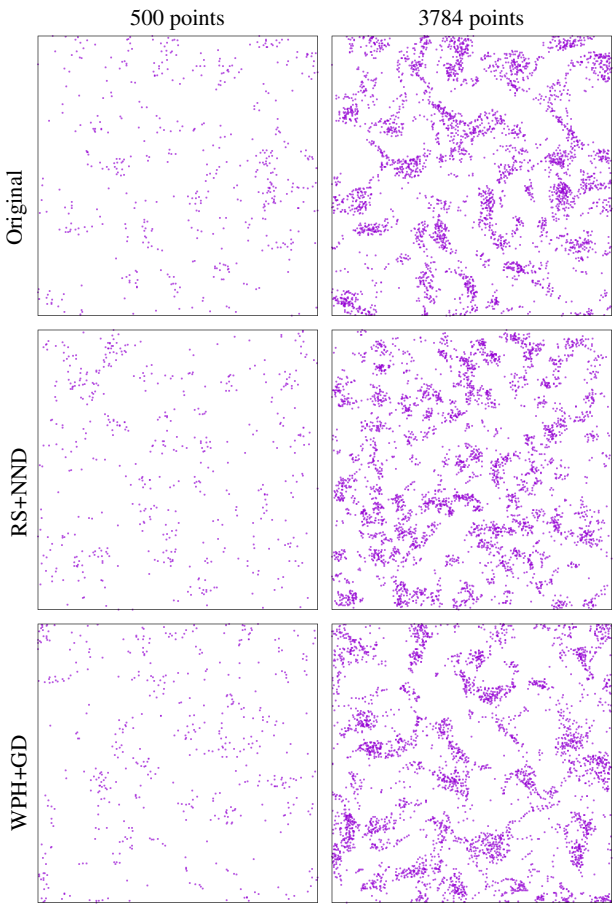

Fig. 7 Syntheses from the RS+NND model and our model, from observations containing 500 points (left) 3784 points (right).

	1.5e-2	2.4e-2	3.3e-2	4.1e-2	5.0e-2
True	5.8e-1(1.1e-2, 1.2e-2)	7.7e-1(1.0e-2, 1.0e-2)	8.9e-1(8.3e-3, 8.5e-3)	9.5e-1(6.3e-3, 6.5e-3)	9.8e-1(4.3e-3, 4.3e-3)
RS+NND	6.0e-1(2.1e-3, 2.1e-3)	8.1e-1(3.9e-3, 4.0e-3)	9.3e-1(4.4e-3, 4.4e-3)	9.8e-1(3.0e-3, 3.1e-3)	1.0e+0(1.5e-3, 1.5e-3)
GD+WPH	5.7e-1(4.2e-3, 4.2e-3)	**7.6e-1** (6.2e-3, 6.4e-3)	**8.9e-1**(6.7e-3, 6.9e-3)	**9.6e-1**(5.6e-3, 5.8e-3)	9.9e-1(3.8e-3, 3.9e-3)

Table 2 Comparison of the SCDF for Turbulence distributions without thinning, on a range of relevant values of r. In parenthesis, the half 95% confidence interval, estimated with bootstrap and standard deviation respectively. Bold numbers indicate values significantly closer to the true distribution.

	8.6e-3	1.6e-2	2.4e-2	3.2e-2	4.0e-2
True	2.7e+1(1.3e+1, 1.3e+1)	-4.1(5.8, 5.7)	-2.7e+1(2.3, 2.4)	-3.1e+1(2.8, 2.9)	-2.3e+1(1.7, 1.7)
RS+NND	9.9e+1(8.2, 8.5)	-1.5e+1(6.1, 5.1)	-4.0e+1(4.0, 4.0)	-4.2e+1(3.1, 3.1)	-2.1e+1(1.4, 1.5)
GD+WPH	**4.6e+1**(1.4e+1, 1.4e+1)	-3.0(4.7, 4.5)	**-2.7e+1**(4.2, 4.2)	-3.2e+1(3.9, 3.9)	-2.9e+1(2.5, 2.5)

Table 3 Comparison of the Euler characteristic for Turbulence distributions without thinning, on a range of relevant values of r. In parenthesis, the half 95% confidence interval, estimated with bootstrap and standard deviation respectively. Bold numbers indicate values significantly closer to the true distribution.

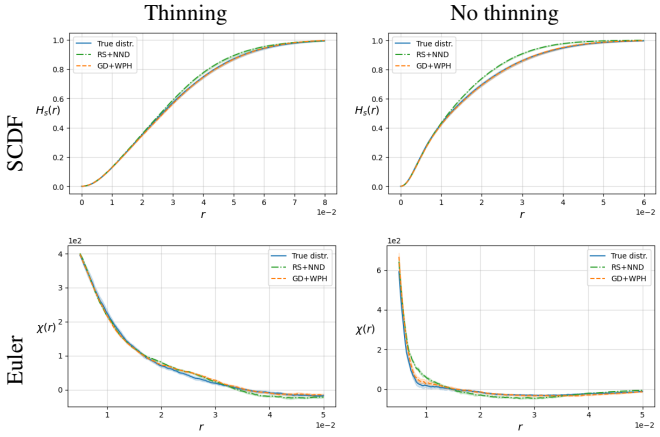

Fig. 8 Spherical contact distribution function (SCDF) and Euler-Poincaré characteristic, for the thinned and not thinned distributions. We compare the statistics of the true distribution, as well as the RS+NND and WPH+GD models.

matrix between (the PD of) all original and synthesized point patterns (cf. Section 6.3), we also compute the average Wasserstein distance between all pairs of point patterns belonging to different distributions. In more details, for a given distribution (Cox Voronoi or Turbulence Poisson), let $M := M_{orig/GD+WPH}$ be the 10×10 distance matrix between the 10 realizations of the original distribution and the 10 realizations of our model. We compute $d_{orig/GD+WPH} = \frac{1}{100}\sum_{i,j} M_{i,j}$. Similarly, we compute $d_{orig/RS+NND}$ for the RS+NND model. We obtained, for the Cox Voronoi example, $d_{orig/GD+WPH} = 0.75$, and $d_{orig/RS+NND} = 1.52$, showing a significant advantage to our method. Our experiments on the Turbulence Poisson example gave $d_{orig/GD+WPH} = 0.61$, and $d_{orig/RS+NND} = 0.62$. These results are coherent with the visual evaluation, that indicates a better performance of our model, particularly for the Cox Voronoi example.

7.3.3 Direct comparison

On the two examples treated above, we shall illustrate the overall performance of both methods. We ran 10 simulations of syntheses from the method of Tscheschel and Stoyan (2006), with $k_{max} = 64$ for the Cox Voronoi example, and $k_{max} = 128$ for the Turbulence Poisson example, with 400 iterations per point. For both models we perform also the syntheses from our multiscale gradient descent method (with 400 iterations).

The two families of examples of syntheses share the same corresponding observation coming from the original distributions. We also simulate 9 other patterns for the original distributions, to compare the averaged statistics and the diversity among the original distribution and the models.

We compare the different distributions with visual evaluation (Figure 9), TDA (Figure 10), and by estimation of the SCDF and Euler characteristic (Figure 11, Tables 4 and 5). For the evaluation with TDA, in addition to the visualization of a 2-dimensional representation of the distance

Figure 11 also confirms our visual evaluation. The curve of the χ function clearly shows a larger error for the RS+NND model. Indeed, we can observe from the alignment of points in the observation that the number of connected components quickly decreases in the patterns of the true distribution. As this alignment is not as well reproduced in the RS+NND model as in ours, we observe that the curve of the χ function decreases more slowly for the RS+NND model. Additionally, the SCDF curves for the Turbulence example show a significant error for the RS+NND model, for which the curve is above the true distribution curve, indicating the presence of fewer large empty areas around clusters. For the Voronoi example however, our model also shows a significant error on the SCDF curve, similar to the RS+NND model, possibly due to the presence of points inside the formed cells (only one is needed to impede the presence of an empty region or a hole). This can also explain the error observe on the TDA plot of the Voronoi distributions (Figure 10, left).

	1.5e-2	2.3e-2	3.0e-2	3.7e-2	4.5e-2
True	5.8e-1(1.1e-2, 1.2e-2)	7.5e-1(1.0e-2, 1.1e-2)	8.6e-1(8.8e-3, 8.9e-3)	9.3e-1(7.1e-3, 7.3e-3)	9.7e-1(5.5e-3, 5.6e-3)
RS+NND	5.9e-1(1.4e-3, 1.4e-3)	7.8e-1(1.8e-3, 1.8e-3)	8.9e-1(2.5e-3, 2.6e-3)	9.6e-1(3.3e-3, 3.3e-3)	9.9e-1(2.9e-3, 3.0e-3)
GD+WPH	5.7e-1(4.2e-3, 4.2e-3)	7.4e-1(6.1e-3, 6.3e-3)	**8.6e-1**(7.0e-3, 7.1e-3)	**9.3e-1**(6.2e-3, 6.3e-3)	9.7e-1(4.8e-3, 4.9e-3)

Table 4 Comparison of the SCDF for Turbulence distributions, on a range of relevant values of r. In parenthesis, the half 95% confidence interval, estimated with bootstrap and standard deviation respectively. Bold numbers indicate values significantly closer to the true distribution.

	1.0e-2	2.0e-2	3.0e-2	4.0e-2	5.0e-2
True	1.5e+2(7.2, 7.4)	-5.3e+1(5.0, 5.2)	-6.4e+1(4.2, 4.3)	-4.4e+1(3.2, 3.2)	-2.3e+1(1.7, 1.7)
RS+NND	2.5e+2(3.9, 3.7)	-6.5(2.1, 2.1)	-5.4e+1(3.3, 3.4)	-3.2e+1(3.0, 3.1)	-1.2e+1(1.2, 1.2)
GD+WPH	**1.8e+2**(3.1, 3.1)	**-4.1e+1**(3.2, 3.3)	-7.1e+1(2.1, 2.2)	-4.6e+1(1.9, 1.9)	-1.3e+1(8.5e-1, 8.6e-1)

Table 5 Comparison of the Euler characteristic for Voronoi distributions, on a range of relevant values of r. In parenthesis, the half 95% confidence interval, estimated with bootstrap and standard deviation respectively. Bold numbers indicate values significantly closer to the true distribution.

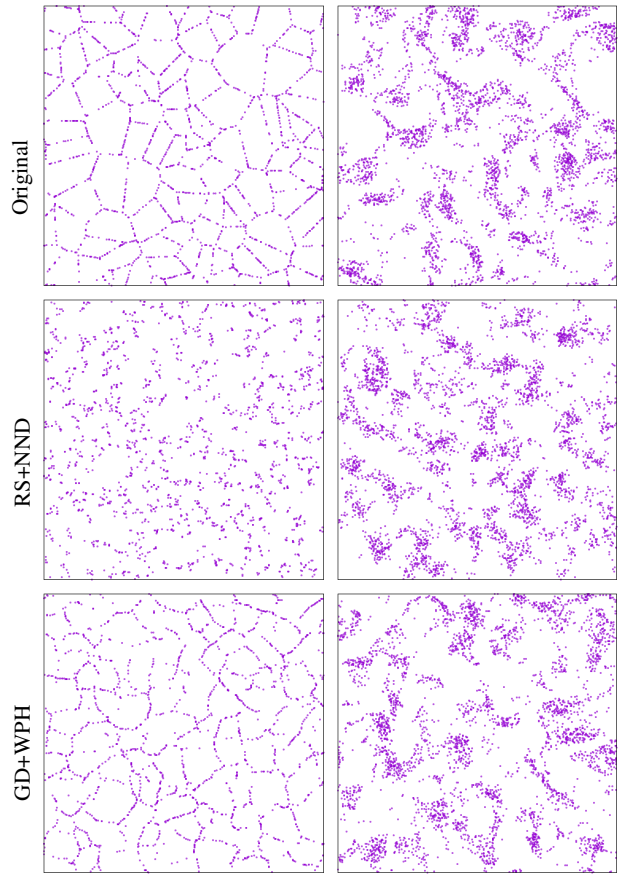

Fig. 9 Visualization of syntheses for the Cox Voronoi and Turbulence Poisson examples. Top: observations, middle: RS+NND, bottom: GD+WPH.

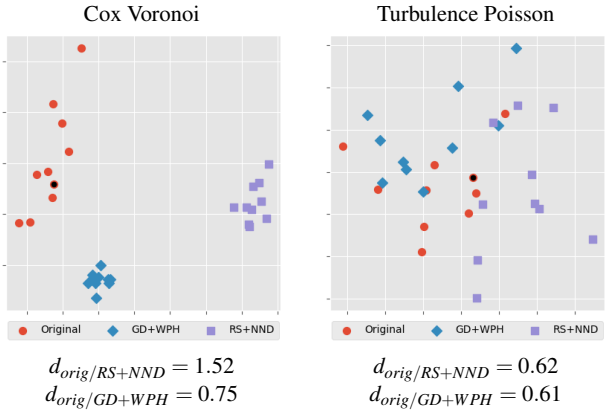

Fig. 10 Visualization of the TDA of the three distributions (Original, GD+WPH, RS+NND), for the Cox Voronoi example (left), and the Turbulence Poisson example (right). The black point represents the observation pattern used for the syntheses. The averaged (true) distances between the original and synthesized patterns (via Wasserstein distance of persistence diagrams) are given as well.

8 Conclusion

In this paper, we present a particle gradient descent model to simulate stationary and ergodic point processes, based on a single observation in a square window. This model is able to synthesize processes formed by a large number of points, exhibiting interactions at multiple scales. Our method is built upon recent works on gradient descent methods to approximate the micro-canonical model. To characterize complex geometric point patterns, we use the wavelet phase harmonic descriptors that allow to explicitly control the scales of the structures to model. Numerical results on Cox and Turbulent distributions validate the ability of the model to capture various geometric structures in the observation. Compared to the classical approaches developed in Torquato (2002); Tscheschel and Stoyan (2006), our approach brought a new perspective to the modeling of point processes, through the lens of wavelet analysis and image modeling.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Baccelli F, Woo JO (2016) On the entropy and mutual information of point processes. In: 2016 IEEE International Symposium on Information Theory (ISIT), IEEE, pp 695–699

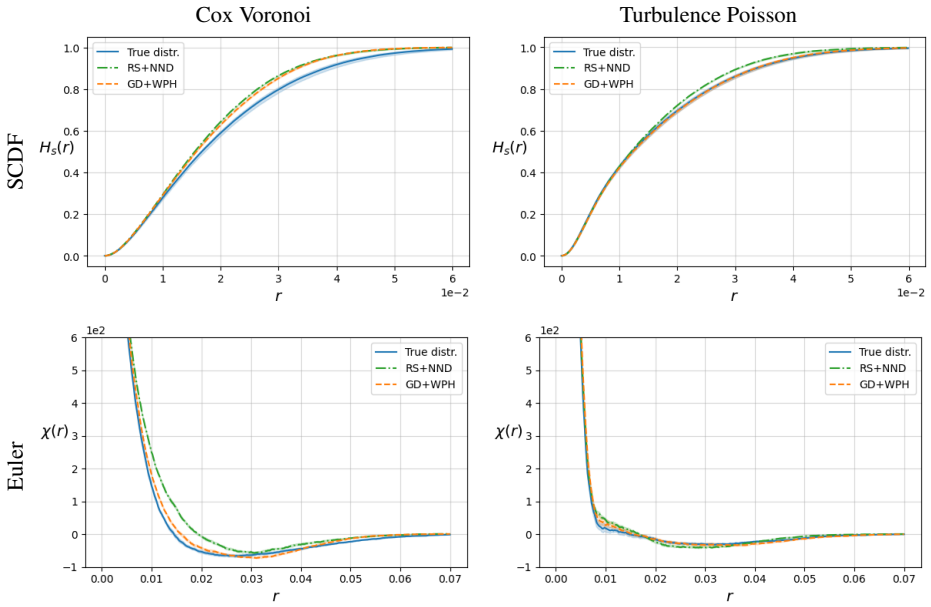

Fig. 11 Spherical contact distribution function (SCDF) and Euler-Poincaré characteristic, for the Voronoi and Turbulence distributions. We compare the statistics of the true distributions, the RS+NND, and the WPH+GD models.

Baddeley A, Jammalamadaka A, Nair G (2014) Multitype point process analysis of spines on the dendrite network of a neuron. Journal of the Royal Statistical Society: Series C: Applied Statistics pp 673–694

Bartlett MS (1964) The spectral analysis of two-dimensional point processes. Biometrika 51(3/4):299–311

Boissonnat JD, Chazal F, Yvinec M (2018) Geometric and topological inference, vol 57. Cambridge University Press

Brémaud P (2002) Mathematical Principles of Signal Processing: Fourier and Wavelet Analysis. Springer Science & Business Media

Brochard A, Błaszczyszyn B, Mallat S, Zhang S (2020) Particle gradient descent model for point process generation. arXiv preprint arXiv:201014928

Brumwell X, Sinz P, Kim KJ, Qi Y, Hirn M (2018) Steerable Wavelet Scattering for 3D Atomic Systems with Application to Li-Si Energy Prediction. arXiv preprint arXiv:181202320

Bruna J, Mallat S (2019) Multiscale sparse microcanonical models. Mathematical Statistics and Learning 1(3):257–315

Chazal F, Michel B (2017) An introduction to topological data analysis: fundamental and practical aspects for data scientists. arXiv preprint arXiv:171004019

Chenouard N, Unser M (2011) 3D steerable wavelets and monogenic analysis for bioimaging. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp 2132–2135

Chiu SN, Stoyan D, Kendall WS, Mecke J (2013) Stochastic geometry and its applications. John Wiley & Sons

Daley DJ, Vere-Jones D (2008) An Introduction to the Theory of Point Processes, vol. II, Probability and Its Applications, vol 2. Springer New York, New York, NY

Dereudre D (2019) Introduction to the theory of gibbs point processes. In: Stochastic Geometry, Springer, pp 181–229

Diggle PJ, Eglen SJ, Troy JB (2006) Modelling the bivariate spatial distribution of amacrine cells. In: Case Studies in Spatial Point Process Modeling, Springer, pp 215–233

Ducasse L, Pumir A (2008) Intermittent particle distribution in synthetic free-surface turbulent flows. Physical Review E 77(6):066304

Ducasse L, Pumir A (2009) Inertial particle collisions in turbulent synthetic flows: quantifying the sling effect. Physical Review E 80(6):066312

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC press

Fasy BT, Kim J, Lecci F, Maria C, Rouvreau V (2014) TDA: statistical tools for topological data analysis. Software available at https://cran r-project org/package= TDA

Gatys L, Ecker AS, Bethge M (2015) Texture synthesis using convolutional neural networks. Advances in neural information processing systems 28:262–270

Geman S, Geman D (1984) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence 6(6):721–741

Illian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistical analysis and modelling of spatial point patterns, vol 70. John Wiley & Sons

Jaynes ET (1957) Information theory and statistical mechanics. Physical review 106(4):620

Koňasová K, Dvořák J (2021) Stochastic reconstruction for inhomogeneous point patterns. Methodology and Computing in Applied Probability 23(2):527–547

Liu DC, Nocedal J (1989) On the limited memory bfgs method for large scale optimization. Mathematical programming 45(1):503–528

Mallat S (2001) A Wavelet Tour of Signal Processing: The Sparse Way, 3rd Edition. Academic Press

Mallat S, Zhang S, Rochette G (2020) Phase harmonic correlations and convolutional neural networks. Information and Inference: A Journal of the IMA 9(3):721–747

Matsuda K, Onishi R (2019) Turbulent enhancement of radar reflectivity factor for polydisperse cloud droplets. Atmospheric Chemistry and Physics 19(3):1785–1799

Molchanov I, Zuyev S (2002) Steepest descent algorithms in a space of measures. Statistics and Computing 12(2):115–123

Müller R, Schuhmacher D, Mateu J (2020) Metrics and barycenters for point pattern data. Statistics and Computing 30(4):953–972

Oujia T, Matsuda K, Schneider K (2020) Divergence and convergence of inertial particles in high-reynolds-number turbulence. Journal of Fluid Mechanics 905

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. (2019) Pytorch: An imperative

style, high-performance deep learning library. Advances in neural information processing systems 32:8026–8037
Portilla J, Simoncelli EP (2000) A parametric texture model based on joint statistics of complex wavelet coefficients. International journal of computer vision 40(1):49–70
Rabin J, Peyré G, Delon J, Bernot M (2011) Wasserstein barycenter and its application to texture mixing. In: International Conference on Scale Space and Variational Methods in Computer Vision, Springer, pp 435–446
Rajala T, Redenbach C, Särkkä A, Sormani M (2018) A review on anisotropy analysis of spatial point patterns. Spatial Statistics 28:141–168
Schneider K, Ziuber J, Farge M, Azzalini A (2006) Coherent vortex extraction and simulation of 2d isotropic turbulence. Journal of Turbulence 7(44):N44
Skare Ø, Møller J, Vedel Jensen EB (2007) Bayesian analysis of spatial point processes in the neighbourhood of voronoi networks. Statistics and Computing 17(4):369–379
Stoica RS, Martinez VJ, Mateu J, Saar E (2005) Detection of cosmic filaments using the candy model. Astronomy & Astrophysics 434(2):423–432
Stoyan D, Stoyan H (1994) Fractals, random shapes and point fields: methods of geometrical statistics, vol 302. Wiley-Blackwell
Tempel E, Stoica RS, Kipper R, Saar E (2016) Bisous model—detecting filamentary patterns in point processes. Astronomy and Computing 16:17–25
Torquato S (2002) Random Heterogeneous Materials: Microstructure and Macroscopic Properties. Springer, New York,
Tscheschel A, Stoyan D (2006) Statistical reconstruction of random point patterns. Computational statistics & data analysis 51(2):859–871
Wadhwa RR, Williamson DF, Dhawan A, Scott JG (2018) TDAstats: R pipeline for computing persistent homology in topological data analysis. Journal of open source software 3(28):860
Wiegand T, Moloney KA (2013) Handbook of spatial point-pattern analysis in ecology. CRC press
Zhang G, Stillinger FH, Torquato S (2015a) Ground states of stealthy hyperuniform potentials: I. entropically favored configurations. Physical Review E 92(2):22119
Zhang G, Stillinger FH, Torquato S (2015b) Ground states of stealthy hyperuniform potentials. ii. stacked-slider phases. Physical Review E 92(2):022120
Zhang S, Mallat S (2021) Maximum entropy models from phase harmonic covariances. Applied and Computational Harmonic Analysis 53:199–230

A Proof of Theorem 1

In order to prove Theorem 1, we need to formally define Eq. (9). Recall that in this section and in what follows, W_s is interpreted as endowed with the addition and scalar multiplication modulo W_s.

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

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

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

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

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

$$\nabla_x K(\mu) := \begin{cases} Jac[K_x^\mu](0) & \text{if } x \in \text{Supp}(\mu) \\ 0 & \text{otherwise} \end{cases}, \tag{18}$$

and

$$\nabla_x E_{\bar\phi}(\mu) := \begin{cases} Jac[E_x^\mu](0) & \text{if } x \in \text{Supp}(\mu) \\ 0 & \text{otherwise} \end{cases}, \tag{19}$$

where $Jac[f]$ denotes the Jacobian matrix of the function f. When $x \in \text{Supp}(\mu)$, the chain-rule gives $Jac[E_x^\mu](0) = (\nabla_x K(\mu))^t (K(\mu) - K(\bar\phi))$. We can now give the proof of Theorem 1.

Proof We are going to show that if Φ_n follows a distribution invariant to T, then $\Phi_{n+1} = G_{\bar\phi}(\Phi_n)$ also follows a distribution that is invariant to T. The gradient descent procedure thus produces a sequences of measures Φ_n that are all invariant to T because the initial random measure Φ_0 is invariant invariant to T.

Denote $G_{\bar\phi}(\mu)$ by the measure configuration transported from μ, by performing one gradient-descent step on the energy $E_{\bar\phi}$. More precisely, for $\mu = \sum_i \delta_{x_i}$, we define for a fixed $\gamma > 0$, the gradient-descent step by

$$G_{\bar\phi}(\mu) := \sum_i \delta_{x_i - \gamma \nabla_{x_i} E_{\bar\phi}(\mu)}.$$

For any transform $Tx = Ax + b$ on W_s, where A is an orthogonal matrix A with entries in $\{-1, 0, 1\}$, and $b \in W_s$. As A is a linear transformation on the torus W_s, $\forall x, y \in W_s$, $A(x+y) = Ax + Ay$. We shall first prove that

$$G_{\bar\phi}(T_\# \mu) = T_\# G_{\bar\phi}(\mu). \tag{20}$$

Let $y_i = Tx_i$, then by definition,

$$G_{\bar\phi}(T_\#\mu) = \sum_i \delta_{y_i - \gamma \nabla_{y_i} E_{\bar\mu}(T_\#\mu)},$$

and

$$T_\# G_{\bar\phi}(\mu) = \sum_i \delta_{T(x_i - \gamma \nabla_{x_i} E_{\bar\phi}(\mu))}.$$

We are going to show that for each i-th particle, $y_i - \gamma \nabla_{y_i} E_{\bar\phi}(T_\#\mu) = T(x_i - \gamma \nabla_{x_i} E_{\bar\phi}(\mu))$. This implies that (20) is correct. The key is to show that

$$A \nabla_{x_i} E_{\bar\phi}(\mu) = \nabla_{y_i} E_{\bar\phi}(T_\#\mu) \tag{21}$$

which will imply that $\forall i$,

$$T(x_i - \gamma \nabla_{x_i} E_{\bar\phi}(\mu)) = Ax_i - \gamma A \nabla_{x_i},$$

$$E_{\bar\phi}(\mu) + b = y_i - \gamma A \nabla_{x_i} E_{\bar\phi}(\mu) = y_i - \gamma \nabla_{y_i} E_{\bar\phi}(T_\#\mu).$$

To show (21), we recall that by the definitions in Section 3.1,

$$\nabla_{x_i} E_{\bar\phi}(\mu) = Jac(K_{x_i}^\mu)(0)^t (K(\mu) - K(\bar\phi)), \tag{22}$$

$$\nabla_{y_i} E_{\bar\phi}(T_\#\mu) = Jac(K_{y_i}^{T_\#\mu})(0)^t (K(T_\#\mu) - K(\bar\phi)), \tag{23}$$

with $Jac(K_{y_i}^{T_\#\mu})(0) = Jac(K_{y_i}^{T_\#\mu} \circ A \circ A^{-1})(0) = Jac(K \circ h_{y_i}^{T_\#\mu} \circ A)(0) A^{-1}$. Furthermore, $\forall x \in W_s$,

$$\begin{aligned} K \circ h_{y_i}^{T_\#\mu} \circ A(x) &= K(T_\#\mu - \delta_{y_i} + \delta_{y_i + Ax}) = K(T_\#(\mu - \delta_{x_i} + \delta_{x_i + x})) \\ &= K(\mu - \delta_{x_i} + \delta_{x_i + x}) = K \circ h_{x_i}^\mu(x), \end{aligned} \tag{24}$$

where we used the fact that T is affine, and the invariance of K w.r.t. T. The equality in (21) follows directly from (22),(23),(24) and the fact that $A^{-1} = A^t$. From (21), we conclude that (20) holds.

Based on (20), it remains to show that for any Borel set Γ on W_s, $P(\Phi_n \in T_\#^{-1}\Gamma) = P(\Phi_n \in \Gamma)$. This can be shown by induction, since by assumption it holds at $n = 0$: assume now that this statement holds at $n \geq 0$, then we have,

$$\begin{aligned} P(\Phi_{n+1} \in T_\#^{-1}\Gamma) &= P(G_{\tilde{\phi}}(\Phi_n) \in T_\#^{-1}\Gamma) \quad (25)\\ &= P(T_\# G_{\tilde{\phi}}(\Phi_n) \in \Gamma)\\ &= P(G_{\tilde{\phi}}(T_\#\Phi_n) \in \Gamma)\\ &= P(G_{\tilde{\phi}}(\Phi_n) \in \Gamma)\\ &= P(\Phi_{n+1} \in \Gamma). \quad (26) \end{aligned}$$

The second last equality in (26) is due to the invariance of Φ_n, i.e.

$$\begin{aligned} P(G_{\tilde{\phi}}(T_\#\Phi_n) \in \Gamma) &= P(\Phi_n \in T_\#^{-1} G_{\tilde{\phi}}^{-1}\Gamma)\\ &= P(\Phi_n \in G_{\tilde{\phi}}^{-1}\Gamma)\\ &= P(G_{\tilde{\phi}}(\Phi_n) \in \Gamma). \end{aligned}$$

B Fourier spectrum and power spectrum

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

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

Observe, $F_m(\mu)$ at frequency $m = (0,0)$ specifies the number of points of the measure μ on W_s. The empirical *Fourier spectrum* (or *power spectrum*) is often defined by taking the square modulus of the Fourier coefficients $F_m(\mu)$; $U_m(\mu) := |F_m(\mu)|^2$. Note that $|F_m(\mu)|^2$, 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. As we shall focus on isotropic point processes, we further reduce the variance of our statistics 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 $\{m \in \Gamma_F : \lfloor |m| \rfloor = k\}$. The radial power spectrum $P(k)$ is the expectation of $\tilde{U}_k(\mu)$ for $k \in N = \{1,2,3,...\}$ when μ follows some distribution, divided by the intensity of the process (estimated over 10 realizations).

C Relaxing the assumptions on the data

In this paper, in order to present our model in a simple setting, strong theoretical assumptions have been made on the data. However, in real world applications, the data will most likely not satisfy these assumptions. This sections presents ideas on how to adapt our model in such cases.

Non-periodic boundaries Recall that our descriptor, defined in (14), applies periodic boundary correction to point patterns in a square window. If the structure of the observed pattern is not periodic, one can modify the descriptor by applying non-periodic integrals in (14) 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 (14) with $i = (\lambda, k, \lambda', k', \tau') \in \Gamma_H$ over smaller windows $W_{s_i} \subset W_s$, such that boundary effects are negligible. Our current software can also handle such non-periodic boundary conditions.

More general observation windows In this paper, we considered that the observed pattern lies in a square observation window. If this is not the case, one could use a similar idea to the non-periodic case: embed the observation window in a square window and considering integrals in (14) over the observation window.

Non stationary process In Koňasová and Dvořák (2021), the authors focus on building a model for non stationary point processes inspired by Tscheschel and Stoyan (2006). Similarly, one might adapt our method to model non stationary processes. This could be done by modifying two aspects of the method. First, the initial distribution Φ_0 (cf. Section 3.1) could be chosen as a non stationary Poisson point process, estimating the intensity with a kernel estimator, such as in Koňasová and Dvořák (2021). In addition, as pointed out in Koňasová and Dvořák (2021), the descriptor should be adapted not to be translation invariant. This could be done, for example, by applying local integrals over patches of the observation window in (14) (this requires some notion of "local stationarity" of the process). Another method that may be useful in this scenario is the regularization proposed in Brochard et al. (2020), where a regularization term is added to the energy. This term consists of the (Sliced Wasserstein) distance (Rabin et al. 2011) between the initial configuration and the current configuration (the one being optimized). By adding this regularization term to the energy, the points of the configuration are forced not to move too far away from the initial configuration, which could help preserve the non stationarity of the initial distribution in the distribution of the model.

Processes in other dimensions While we focus in this paper on planar point processes, our approach can readily be extended to any dimensions. To model point processes in other dimensions such as 1d or 3d, one can consider similar type of wavelets proposed in the literature (Chenouard and Unser 2011; Brumwell et al. 2018).

D List of important parameters of our model

In Table 6, we discuss the main parameters of our model in three categories. The first two categories are the parameters that are relatively standard to consider in most existing methods such as Tscheschel and Stoyan (2006). The third category is more specific to our model, which involves the discretization step, and the final blurring step.

Category	Parameter	Discussion
Descriptor (c.f. Section 4.3 and 5.1.2)	Number of wavelet scales J	The scales of the wavelet transform are defined by $0 \leq j < J$. The minimal scale $j = 0$ is chosen though the image resolution N of the discretization, and is related to the precision in high frequencies. This choice should depend on the observed pattern. The maximal scale J should be as large as possible, to capture enough structural information, but not too large, so that the wavelet phase harmonic covariances remain empirically well estimated.
	Number of wavelet orientations L	This determines the angular precision of the descriptor. A larger L captures finer orientations of edge-like structures.
	Range of phase harmonics (k,k')	This determines the range of interactions between the wavelet phase harmonics coefficients. The choice of $(k,k') = (1,1)$ corresponds to the second order statistics.
Optimisation (cf Section 3.1 and 6.1.2)	Number of iterations	In our experiments, we set a fixed number of iterations. We found that increasing the number of iterations further only decrease the energy of the configurations by a small factor. Other standard stopping criteria based on the norm of gradient can also be considered.
Extra steps (cf. Section 5)	Image resolution N	The larger the resolution, the smaller the structures of point processes which we can model. However, there is an extra computational cost when N increases. It also results in a larger number of moments to estimate.
	With or without multi-scale optimization	Multi-scale optimization has been found useful in the case where the maximal scale J is large, to avoid poor local minima and reconstruct the observation. We have also used it in our synthesis experiments, to reduce the energy of the syntheses.
	Final blurring	The final blurring is useful in the cases where the number of points per pixel is often larger than 1, to remove artifacts due to the discretization.

Table 6 Discussion of the important parameters of our model.