Finding wombling boundaries in LHC data with Voronoi and Delaunay tessellations

Konstantin T. Matchev, Alexander Roman and Prasanth Shyamsundar

Institute for Fundamental Theory, Physics Department, University of Florida,
Gainesville, FL 32611, U.S.A.

E-mail: matchev@ufl.edu, alexroman@ufl.edu, prasanths@ufl.edu

ABSTRACT: We address the problem of finding a wombling boundary in point data generated by a general Poisson point process, a specific example of which is an LHC event sample distributed in the phase space of a final state signature, with the wombling boundary created by some new physics. We discuss the use of Voronoi and Delaunay tessellations of the point data for estimating the local gradients and investigate methods for sharpening the boundaries by reducing the statistical noise. The outcome from traditional wombling algorithms is a set of boundary cell candidates with relatively large gradients, whose spatial properties must then be scrutinized in order to construct the boundary and evaluate its significance. Here we propose an alternative approach where we simultaneously form and evaluate the significance of all possible boundaries in terms of the total gradient flux. We illustrate our method with several toy examples of both straight and curved boundaries with varying amounts of signal present in the data.

Dedicated to B.P. Delaunay.

Contents

1 Introduction

Collider data in high energy physics can be viewed, at least at the parton level, as a collection of points $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$ in the relevant phase space $\mathcal{P}$ of the final state signature. The ultimate goal of a high-energy physics experiment is then to test whether the distribution of those N points (commonly referred to as "events") in $\mathcal{P}$ follows the probability distribution predicted in some theory model by the fully differential cross-section

$$\frac{d\sigma}{d\vec{x}} \equiv f(\vec{x}, \{\alpha\}), \tag{1.1}$$

where $\vec{x} \in \mathcal{P}$ is a particular phase space point and $\{\alpha\}$ is a set of input model parameters such as particle masses, widths, couplings, etc.. More specifically, in searches for new physics (NP) beyond the standard model (SM), eq. (1.1) can be split into respective SM and NP contributions

$$f(\vec{x}, \{\alpha\}) \equiv f_{\mathrm{SM}}(\vec{x}, \{\alpha_{\mathrm{SM}}\}) + f_{\mathrm{NP}}(\vec{x}, \{\alpha_{\mathrm{SM}}\}, \{\alpha_{\mathrm{NP}}\}), \tag{1.2}$$

where $\{\alpha_{\mathrm{SM}}\}$ and $\{\alpha_{\mathrm{NP}}\}$ label respectively the set of SM parameters and the set of additional NP parameters. Let the corresponding regions of phase space populated by SM and NP events be $\mathcal{P}_{\mathrm{SM}}$ and $\mathcal{P}_{\mathrm{NP}}$, then the respective total cross-sections are given by

$$\sigma_{\mathrm{SM}} = \int_{\mathcal{P}_{\mathrm{SM}}} f_{\mathrm{SM}}(\vec{x}, \{\alpha_{\mathrm{SM}}\}) \, d\vec{x}, \tag{1.3a}$$

$$\sigma_{\mathrm{NP}} = \int_{\mathcal{P}_{\mathrm{NP}}} f_{\mathrm{NP}}(\vec{x}, \{\alpha_{\mathrm{SM}}\}, \{\alpha_{\mathrm{NP}}\}) \, d\vec{x}. \tag{1.3b}$$

Since the standard model is well known, its distribution f_{SM}, a.k.a. "the background", is calculable theoretically (up to some fixed order in perturbation theory). However, once we account for the experimental realities, the result may suffer from non-negligible systematic uncertainties, particularly in the case of challenging signatures involving QCD and/or reducible backgrounds. This is the main roadblock in NP searches via counting experiments, where one focuses on a suitably chosen "signal region" $\mathcal{P}_{SR} \subset \mathcal{P}_{\mathrm{NP}}$ and looks for an excess over the SM expectation $\int_{\mathcal{P}_{SR}} f_{\mathrm{SM}}(\vec{x}, \{\alpha_{\mathrm{SM}}\}) \, d\vec{x}$.[1]

Instead, in this paper we shall consider methods which could allow us to infer, at least in principle, the existence of the NP contribution f_{NP} *without any prior knowledge* of the SM prediction f_{SM}. Recently there has been hightened interest in such "blind" or "background-independent" searches for NP, particularly using machine learning techniques [1–15]. Here, instead of looking for an excess in the signal region $\mathcal{P}_{SR}$, we shall follow up on the idea of refs. [16, 17] to target directly the *boundary* $\partial\mathcal{P}_{\mathrm{NP}}$ of the NP phase space region $\mathcal{P}_{\mathrm{NP}}$, by using the fact that the combined distribution (1.2) is non-differentiable anywhere on $\partial\mathcal{P}_{\mathrm{NP}}$ where f_{NP} is non-vanishing. As it turns out, the latter is a very safe assumption — if anything, f_{NP} is not only non-vanishing, but often enhanced

---

[1]In this paper, we have in mind the typical NP scenarios where $\mathcal{P}_{\mathrm{NP}}$ is a subset of $\mathcal{P}_{\mathrm{SM}}$, $\mathcal{P}_{\mathrm{NP}} \subseteq \mathcal{P}_{\mathrm{SM}}$. This is certainly the case for signatures where $\mathcal{P}_{\mathrm{SM}}$ consists of the full phase space, $\mathcal{P}_{\mathrm{SM}} = \mathcal{P}$. If by any chance the signature happens to be such that $\mathcal{P}_{\mathrm{SM}} \subset \mathcal{P}_{\mathrm{NP}}$, the selection of the signal region is trivial: $\mathcal{P}_{SR} = \mathcal{P}_{\mathrm{NP}} - (\mathcal{P}_{\mathrm{NP}} \cap \mathcal{P}_{\mathrm{SM}})$, and a counting experiment should already be good enough.

and even *singular* on the boundary $\partial \mathcal{P}_{NP}$ [18–27]. Since the background distribution f_{SM} is a smooth function across $\partial \mathcal{P}_{NP}$, the presence of f_{NP} creates a discontinuous "jump" in the combined event density (1.2), precisely at the location of the boundary $\partial \mathcal{P}_{NP}$ [16, 17]. We can thus reformulate the original problem of finding evidence of NP in the collider data as follows:

> Given a collection of N points $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$ in the phase space $\mathcal{P}$, identify (the locations of) any candidate wombling boundaries and estimate their statistical significance.

The detection of such difference boundaries (or wombling boundaries, named after a pioneer in the field, W. H. Womble [28]) is a well-known problem in spatial statistics, see, e.g., [29]. Broadly speaking, wombling is any of a number of techniques used for identifying zones of rapid change, typically in some quantity as it varies across some geographical or Euclidean space. Wombling techniques are being applied in a wide variety of disciplines, including computational ecology, anthropology, linguistics, geography and many others.[2] In our case here, we shall be interested in identifying the phase space region in the vicinity of $\partial \mathcal{P}_{NP}$ where the *density* of points is changing significantly.

Before going through the typical steps of a wombling analysis, several comments are in order. First, the discontinuous jump of the combined distribution (1.2) across $\partial \mathcal{P}_{NP}$ in practice will be smoothed out to some extent by the detector resolution and finite width effects, leading to a well-defined, finite, density gradient everywhere in $\mathcal{P}$. This precludes us from using methods specifically designed to detect image discontinuities such as ridges and cliffs but do not admit gradients [40]. Second, a good wombling method should also be able to pick up any boundaries created within $\mathcal{P}_{SM}$ by interesting SM subprocesses, e.g., top, Higgs or heavy gauge boson production. This would guarantee an opportunity for the LHC experiments to start testing and validating the method with existing real data, before any NP discovery. Third, while most applications of wombling in the literature have been limited to two-dimensional data, the method shoud be readily generalizable to higher dimensions, if it is to be of any interest to the high-energy physics community where the dimensionality of the relevant phase space $\mathcal{P}$ is typically much higher (although in some special cases it can be reduced to 2 or even 1 through suitable projections preserving the boundary). Along those lines, it is also important to choose a good parametrization of $\mathcal{P}$, so that the dimensionality can be reduced by projecting out uninteresting degrees of freedom without washing out the wombling boundary.

The main steps of a typical wombling analysis are the following [41].

- *Data preparation and preprocessing.* The starting point in wombling is a spatially referenced dataset

$$(\vec{x}_i, f_i), \quad i = 1, 2, \ldots, N, \tag{1.4}$$

---

[2]For example, wombling has been used to identify genetic boundaries in Eurasian human populations [30, 31], language boundaries in Europe [32], transition zones in genetic, morphometric and physiological characteristics [33], boundaries of different types of vegetation [34, 35], hospital admission rates for respiratory conditions [36], cancer rates [37], metal concentrations in the Swiss Jura [38] and other environmental data [39].

where a set of values $\{f_i\}$ for the function of interest $f(\vec{x})$ are obtained at some finite number N of point locations $\{\vec{x}_i\}$. In some applications, e.g., for aerial and remotely sensed images, the locations $\{\vec{x}_i\}$ can be chosen by the experimenter — then it may be convenient to arrange them in some kind of a regular lattice, as required by some wombling algorithms, including the original proposal in [28]. Alternatively, when the data are gathered by an irregular or random design, eq. (1.4) is known as *point-referenced* or *geostatistical* data. While in many other fields of science the functional values $\{f_i\}$ for geostatistical data are obtained directly from field observations, for the Monte Carlo simulations used in high-energy physics the situation is more subtle — each f_i is supposed to be a measure of the local point density at $\vec{x}_i$ and needs to be evaluated as a preprocessing step. For this purpose, refs. [16, 17] proposed to consider the Voronoi tessellation in $\mathcal{P}$ of the dataset (1.4), since the geometric volume v_i of the Voronoi cell containing $\vec{x}_i$ provides a natural local estimator of the point density at $\vec{x}_i$ [42]:

$$f_i \sim \frac{1}{v_i}. \tag{1.5}$$

While Voronoi tessellations have been widely used in many other fields of science, they seem to be underutilized in high energy physics where their application has been limited to jet clustering [43] and the partitioning of the signature phase space into search regions, as implemented in the SLEUTH algorithm [44–46] which was used to perform model-independent new physics searches at D0 [47–49], HERA [50] and CDF [51–53]. Yet, the Voronoi approach[3] is ideally suited for finding interesting (e.g., singular) features in $f(\vec{x})$, since it preserves the maximum spatial resolution in the data [54]. For example, the standard approach of binning the data in order to obtain a local density estimate necessarily throws away a certain amount of useful information, and is associated with some arbitrariness in the exact choice of binning [16].

- *Gradient estimation.* Once we have the point-referenced dataset (1.4), the next step is usually to estimate the magnitude of the local gradient $\vec{\nabla} f$ of the function $f(\vec{x})$, since any zone of rapid change is necessarily associated with large values for $|\vec{\nabla} f|$. In calculating the gradient, one has to overcome the fact that the data is a) discrete and b) irregularly sampled. One possible approach is to obtain a continuous approximation for $f(\vec{x})$ via some spatial interpolation method [42, 55]. Unfortunately, interpolation techniques tend to smooth out not only the noise but also small local discontinuities, which can result in masking some true boundaries [34]. For this reason, refs. [16, 17] explored several boundary detection techniques (further developed and illustrated in [22, 23, 56]) which continued to use the Voronoi tessellation of the data and the fundamental relation (1.5). Among the different options studied in [16, 56], the normalized standard deviation (sometimes also called the coefficient of variance) of the volumes of the neighboring cells emerged as a viable measure of

---

[3]In the context of treating high energy collider data as a point dataset (1.4), it is worth mentioning the recent idea of ref. [14] to consider $\{\vec{x}_i\}$ as a graph network of weighted nodes, which is somewhat orthogonal, but similar in spirit to the Voronoi approach.

the magnitude of the local gradient within a given Voronoi cell. In this paper, we shall pursue a somewhat orthogonal and more traditional approach, known in the literature as "triangulation wombling" [34, 39], which makes use of the dual Delaunay tessellation of the data (1.4). Even though the two types of tessellations are dual to each other,[4] the Delaunay version seems more natural for the specific problem at hand of calculating gradients — this will be further discussed and illustrated in section 2 below.

- *Tagging selection.* Having constructed the Delaunay or Voronoi tessellation and obtained estimates for the local gradients, the next task is to identify elements of the tessellation (edges or vertices) which are likely to be located on or near a wombling boundary. This is typically accomplished with a cut on the ranked values of $|\vec{\nabla}f|$ for all elements in the tessellation, e.g., selecting elements whose calculated gradients are in a certain upper percentile.[5] However, using such a simple threshold for tagging boundary elements has been viewed as somewhat arbitrary and rather subjective [41] — in the absence of any robust guidelines, typical values used in the literature range in the 5th to 10th percentile [30, 31, 36, 39]. Furthermore, for any given value of the threshold, there will always be a certain number of elements passing the cut, including elements "in the bulk", i.e., away from any wombling boundaries, and one has to design a prescription on how to deal with such false positives. Obviously, a value for the cut which is too stringent will miss many true boundary elements, while a value which is too generous will bring about a lot of false positives. Finally, given that the dimensionalities of $\mathcal{P}_{NP}$ and $\partial\mathcal{P}_{NP}$ necessarily differ by one, the concept of a "boundary element" can be open to interpretation — how close to the boundary does an element have to be in order to be considered a "boundary element"?

- *Agglomeration.* The previous step results in a collection of tagged boundary elements scattered throughout $\mathcal{P}$, so now the question is how to use that information to reconstruct the complete boundary. As a first step, one can start forming sub-boundaries by linking adjacent tagged boundary elements, possibly subject to some additional criteria, e.g., that the directions[6] of their gradients are within $30°$ of each other [31]. This agglomeration procedure will result in a graph whose nodes are the tagged boundary elements [57]. The properties of this graph can then be studied to determine its statistical significance [32, 34] (see the next item) and to get some idea about the shape of the boundary.

---

[4]In graph theory, the dual graph of a plane graph G is a graph that has a vertex for each face of G. Correspondingly, each edge e of G has a corresponding dual edge, whose endpoints are the dual vertices corresponding to the faces on either side of e. Some care must be exercised for the Delaunay edges along the convex hull of the point set — their Voronoi duals are infinite rays which can be turned into finite line segments by adding an artificial point at infinity which serves as a common endpoint for all the rays. In our analysis, this complication will not arise since we will perform our analysis in the interior of $\mathcal{P}$.

[5]In NP scenarios where the signal density is additionally *enhanced* on the boundary $\partial\mathcal{P}_{NP}$, ref. [22] proposed a two-dimensional cut, simultaneously targeting both large gradients and large values of the function.

[6]The two requirements — that the gradients are large *and* that their directions are correlated — can be conveniently encoded in the scalar (dot) product of the gradient vectors of neighboring elements [16].

At this point it is worth mentioning that in collider physics applications there can be situations where the shape of the boundary is parametrically known. This is precisely the case with "simplified model" NP searches at the LHC [58] — once the event topology is assumed, the geometry of the NP final state phase space $\mathcal{P}_{\mathrm{NP}}$ is also fixed. As an example, consider a sequence of three two-body decays, which is the classic squark signature in supersymmetry (SUSY) [59]. The relevant phase space $\mathcal{P}_{\mathrm{NP}}$ is three-dimensional and can be parametrized by the invariant masses of the three pairs of visible decay products. The equation for the boundary $\partial\mathcal{P}_{\mathrm{NP}}$ is known analytically [21, 60, 61] in terms of just four parameters — the masses of the SUSY particles participating in the decay chain. In that situation, ref. [23] proposed to bypass *both* of the last two steps (the tagging and the agglomeration) altogether and instead fit the equation for the surface $\partial\mathcal{P}_{\mathrm{NP}}$ to the full tessellation. Operationally this was done by computing a quantity inspired by Bayesian wombling (see next bullet), namely, a two-dimensional surface integral of the gradient magnitude over the boundary surface, normalized to the total area of the surface:

$$\frac{\int_{\partial\mathcal{P}_{\mathrm{NP}}} da \ |\vec{\nabla} f|}{\int_{\partial\mathcal{P}_{\mathrm{NP}}} da}. \tag{1.6}$$

In ref. [23], it was demonstrated that this quantity is maximized for the true values of the SUSY masses, resulting in a novel method for SUSY mass measurements.

- *Significance estimation.* Any wombling algorithm as described so far will produce numerical results regardless of whether a true pattern exists or not. The crucial question now is to assess the likelihood that the observed pattern could have been produced from random fluctuations instead of a true boundary. One possible approach, known as sub-boundary statistics [32, 34], is to analyze the properties of the graph mentioned in the previous step, formed out of the tagged boundary elements. Strictly speaking, sub-boundary statistics tests whether the different components of the graph are sufficiently contiguous (and not whether the rates of change are sufficiently large) [29, 41]. To this end, one looks at (distributions of) quantities which would characterize coherent boundaries formed from connected boundary elements, such as: the total number of subgraphs, the number of single-node subgraphs, the maximum and the mean of the length and/or the diameter of the subgraphs, the superfluity, etc.. An alternative approach, named Bayesian wombling, starts with an ansatz for the shape of the wombling line (in two spatial dimensions) and then computes the average flux of the two-dimensional gradient field through all possible such lines [40, 62, 63]. The idea is that the average flux will be maximized when the ansatz matches the true wombling boundary. As already mentioned in relation to eq. (1.6), the advantage of this approach is that it avoids the subjectiveness associated with the steps of tagging and agglomeration. With either approach, one has to specify a null hypothesis, in order to quantify the confidence level. Unlike other fields of science, where the null hypothesis may not be immediately obvious and one

typically has to rely on a randomization scheme [29], in high-energy physics the null hypothesis is well defined — it is the SM.

In this paper we further develop and refine the Voronoi boundary detection methods from refs. [16, 22, 23, 56]. As before, the main goal will be to outline a method for discovering new physics in LHC collider data by identifying wombling boundaries in phase space. The paper is structured around the five typical steps of algorithmic wombling described above. The novel elements in the analysis presented here are the following:

- In addition to the Voronoi tessellations utilized in [16, 22, 23, 56], here we also consider Delaunay tessellations. In section 2 we shall briefly review the two types of tessellations and outline the range of new possibilities offered by the use of a Delaunay tessellation for our purposes. In particular, in section 2.3 we shall illustrate the four possible types of boundary elements, discuss their relations to each other, and how each can be potentially targeted in the tagging step of the wombling algorithm.

- Unlike previous work in high-energy physics, here our main tool for calculating the local gradients will be the Delaunay tessellation (often referred to as "triangulation" since in two dimensions the Delaunay polygons are triangles). Section 3 is devoted to the topic of gradient estimation — after a brief review in section 3.1 of previous work on estimating gradients from a Voronoi tessellation, in section 3.2 we shall describe the gradient calculation from the Delaunay triangulation. In the process, we shall pay special attention to techniques for reducing the random fluctuations in the obtained gradient values — three such procedures and their interplay and optimization are discussed in section 3.3.

- Different techniques for tagging boundary elements will be discussed in section 4. In addition to tagging Voronoi cells [16, 22], here we shall also be interested in tagging Voronoi edges, Delaunay cells and Delaunay edges as well. Although this is not our main focus here, in section 5 we shall illustrate how these tagging methods can be used for agglomeration.

- The main results of the paper are presented in sections 6–8. We follow the approach of Bayesian wombling [40, 62, 63], which lets us avoid the intricacies and uncertainties of the tagging and agglomeration steps. To gain some intuition, in section 6 we first go over a toy example where the function $f(\vec{x})$ can be sampled continuously from a distribution which resembles real data including finite width effects and detector resolution. Then in section 7 we study point-referenced data of the type (1.4) with a straight (section 7.1) or circular (section 7.2) boundary. The corresponding estimates of the statistical significance of the obtained wombling boundaries are performed in section 8. For simplicity, the illustrative examples in the main body of the paper use data generated from uniform background distributions. For completeness, in appendix A we also present results for two additional examples in which the background distribution is not uniform, but varies according to a power law (section A.1) or an exponential (section A.2).

2 Voronoi and Delaunay tessellations of point data

2.1 Simulation details

Virtually all wombling studies in the literature have been concerned with *two-dimensional* point data generated, e.g., from field samples taken within a certain geographical area, from remotely sensed images, etc.. In this paper, we shall continue to work in two-dimensions, but this will be done only for clarity of the presentation, since it is difficult to visualize Voronoi and Delaunay tessellations in more than two dimensions; the methods which we shall describe will be applicable to higher dimensional data as well. To be specific, we shall consider the Cartesian plane where the data points are specified by their coordinates (x_i, y_i), so that the dataset (1.4) reduces to

$$(x_i, y_i, f_i), \quad i = 1, 2, \ldots, N. \tag{2.1}$$

For concreteness, we shall choose our field of view to be the unit square, $0 \leq x \leq 1$, $0 \leq y \leq 1$, although data will be generated beyond the boundaries of the unit square — this will eliminate any spurious boundary effects like clipping which would modify the statistical properties of the Voronoi cells near the boundaries [64]. Following refs. [16, 17, 22], the datasets will be generated according to (1.2) with the following assumptions for f_{SM} and f_{NP}:

- *Background.* As in previous work [16, 17, 22], our proxy for the SM background distribution f_{SM} will be the uniform distribution[7]

$$f_{SM} = constant. \tag{2.2}$$

  The exact value of the constant will depend on the normalization: for pure-background samples within the unit square the constant is 1, while for background plus signal samples, it will depend on the relative strength of the signal. Strictly speaking, the assumption (2.2) is unrealistic from the point of view of a high energy physicist, since f_{SM} is in general a function of the kinematic variables parametrizing the phase space $\mathcal{P}$. Nevertheless, it is good enough for our purposes here — the important point is that any realistic SM distribution is *very weakly varying* across the boundary $\partial \mathcal{P}_{NP}$, which justifies the use of (2.2) for our model-independent toy examples below. A typical background distribution of $N = 500$ points within the unit square is shown in the left panels of figures 1 and 2. It is evident that such a distribution does not have any obvious features and any wombling boundary would have to be created purely by chance.

- *Signal with a flat boundary.* As our first example of a hypothesized NP signal we shall consider a distribution f_{NP} populating a region $\mathcal{P}_{NP}$ with a flat boundary. Again following refs. [16, 17, 22], we shall take the boundary to be the vertical line at

---

[7]Other choices for the background distribution will be considered in appendix A.
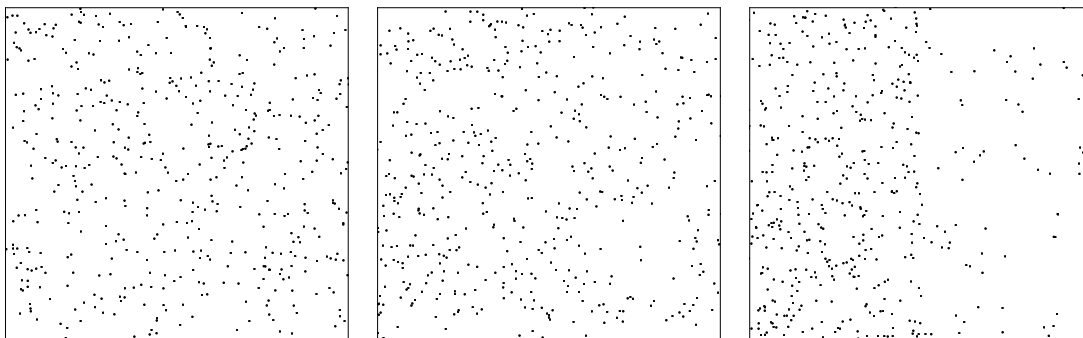
Figure 1. Typical simulated point data sets used in our studies of straight-line boundaries. Here we show $N = 500$ points within the unit square, distributed according to: background only (left panel); background with an additionally injected signal (2.3) in the left half-plane with $\rho = 1.5$ (middle panel) or $\rho = 5$ (right panel).

$x = 0.5$ and the corresponding signal distribution f_{line} to be flat and non-zero only to the left of the boundary:

$$f_{\text{line}} = 2H(0.5 - x), \tag{2.3}$$

where H is the Heaviside step function. When adding this signal to the background distribution, it is important to specify the mixing ratio. For this purpose, refs. [16, 17, 22] introduced a parameter ρ which measures the ratio of the event densities $\frac{dn}{da}$ on the two sides of the boundary:[8]

$$\rho \equiv \frac{\lim\limits_{x \to 0.5^-} \left(\dfrac{dn}{da} \right)}{\lim\limits_{x \to 0.5^+} \left(\dfrac{dn}{da} \right)}. \tag{2.4}$$

Combining (2.2) and (2.3), the unit-normalized total distribution (1.2) on the unit square[9] for a signal with a flat boundary reads [16]

$$f = \frac{2}{\rho + 1} \Big[\rho\, H(0.5 - x) + H(x - 0.5) \Big]. \tag{2.5}$$

In the middle and right panel of figure 1 we show distributions of $N = 500$ points according to (2.5) with $\rho = 1.5$ and $\rho = 5$, respectively. In the latter case, the value of ρ is sufficiently large that the boundary is clearly visible with the naked eye. As in refs. [16, 17, 22] (which considered an even more extreme value of $\rho = 6$), the case of relatively large ρ is meant mostly for illustration — it makes it easier to visualize the benefits from the various wombling and denoising techniques introduced below. Our real target will be the case of relatively low values of ρ as shown in the middle panel of figure 1, where it is rather difficult to discern any apparent wombling boundary.

---

[8]Since we are interested in detecting a wombling boundary, the parameter ρ as defined here is more suitable than the more familiar ratio S/B of signal to background inside the signal region $\mathcal{P}_{SR}$. Note that with our setup, the two are related as $S/B = \rho - 1$.

[9]In order to declutter the notation, in what follows we shall omit the prefactor of $H(x)H(1-x)H(y)H(1-y)$ which confines us to the unit square.
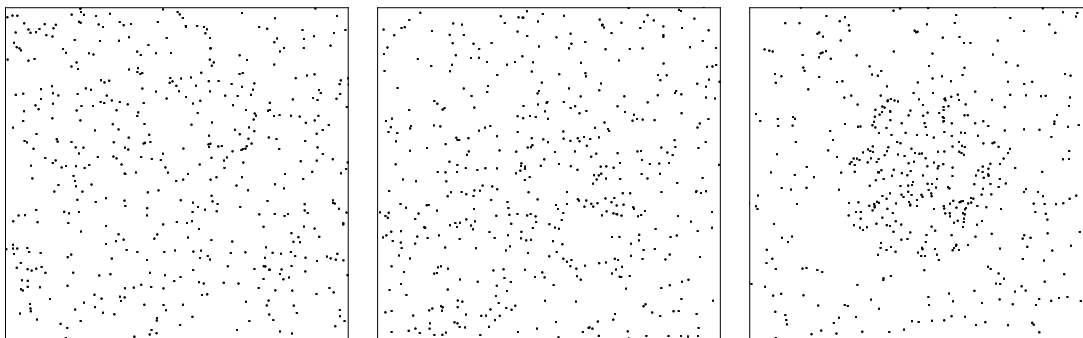
Figure 2. The same as figure 1, but for the circular signal (2.6) with $R = 0.25$.

- *Signal with a circular boundary.* For completeness, we shall also consider an example of a signal in a domain bounded by a curvilinear boundary. Following [22], we shall take the signal distribution f_{circle} to be confined to a circular region of radius $R < 0.5$ centered at $(x, y) = (0.5, 0.5)$:

$$f_{\text{circle}} = \frac{1}{\pi R^2} H\Big(R - \sqrt{(x - 0.5)^2 + (y - 0.5)^2}\Big), \tag{2.6}$$

so that the combined total distribution (1.2) becomes

$$f = \frac{1}{(\rho - 1)\pi R^2 + 1}\Big[\rho\, H\Big(R - \sqrt{(x - 0.5)^2 + (y - 0.5)^2}\Big)$$
$$+ H\Big(\sqrt{(x - 0.5)^2 + (y - 0.5)^2} - R\Big)\Big], \tag{2.7}$$

with the ρ parameter suitably defined in analogy to (2.4) as the ratio of point densities across the circular boundary. The middle and right plots in figure 2 depict typical distributions of $N = 500$ points according to (2.5) with $\rho = 1.5$ and $\rho = 5$, respectively. As before, the boundary for $\rho = 5$ is clearly visible, but the case of $\rho = 1.5$ appears much more challenging.

2.2 Voronoi and Delaunay tessellations

The Voronoi and Delaunay tessellations of the point data sample in the right panel of figure 1 are illustrated in figure 3. In the left panels, which show the Voronoi tessellation, the data points appear as dots, while in the right panels, which illustrate the Delaunay tessellation, the data points are located at the vertices of the Delaunay triangles and are not explicitly shown.

As illustrated in the left panels of figure 3, a Voronoi tessellation of N points (often referred to as *generators*) in the plane is constructed as follows (see, e.g., [42]). Every location in the plane is assigned to the closest member of the point set. If a location happens to be equally close to two (or more) generator points, it is assigned to all of those points; all such eqidistant locations form the *edges* of the Voronoi graph. The set of locations assigned to a given member of the point set forms the Voronoi *cell* corresponding to that generator point; as seen in figure 3, the Voronoi cells in the plane are polygons,
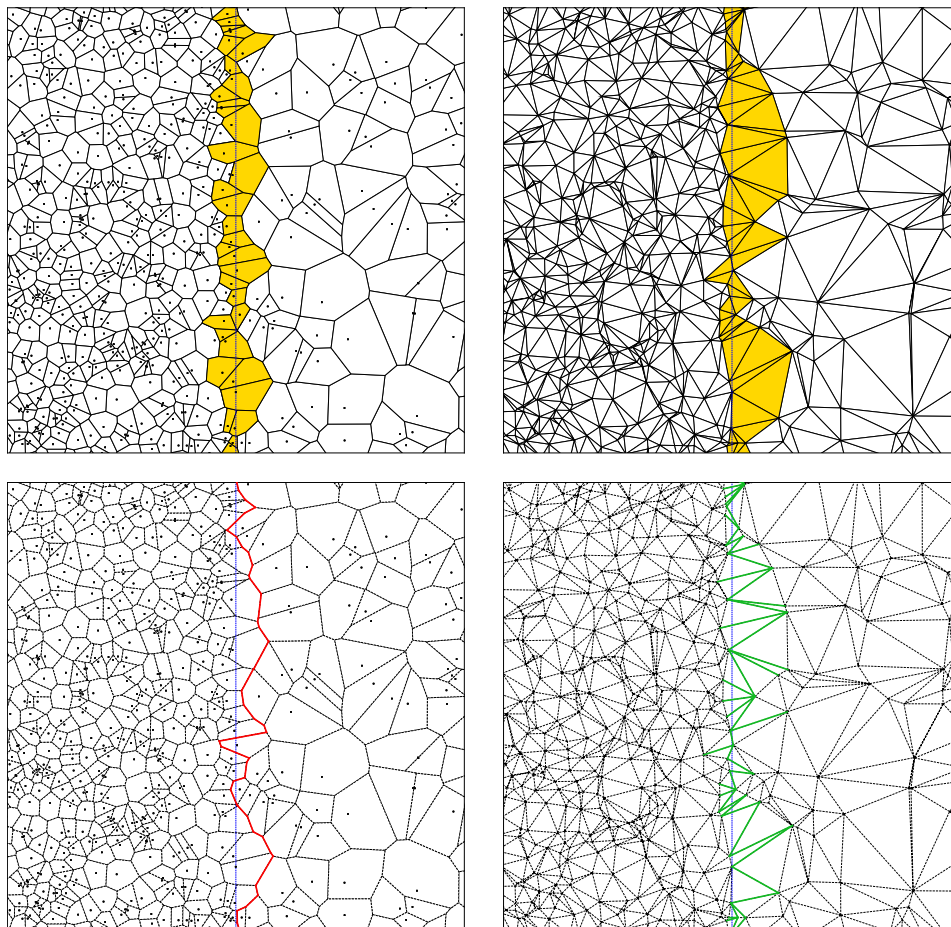
Figure 3. The Voronoi and Delaunay tessellations of a typical point data distributed according to eq. (2.5) with $N = 500$ and $\rho = 4$. The left (right) panels show the Voronoi (Delaunay) tessellation. The vertical blue dashed line marks the theoretical boundary line at $x = 0.5$. The yellow-shaded cells in the top panels illustrate the boundary cells defined in section 2.3 for the Voronoi and Delaunay case, respectively. In the bottom panels, the tessellations are outlined with dotted lines while the solid lines represent the boundary Voronoi edges (red lines) and their dual Delaunay edges (green lines) as defined in section 2.3.

with the corresponding generator point located somewhere in the polygon's interior, but not necessarily at its geometric center. Note that Voronoi polygons have different shapes and sizes; in particular, the number of edges of a polygon varies greatly, the average number being no more than six [42]. An endpoint of a Voronoi edge is called a Voronoi *vertex*; alternatively, a vertex may be defined as a point shared by three (or more) Voronoi edges. When each vertex belongs to three and only three edges, the Voronoi tessellation is non-degenerate; as seen in figure 3 this will be our case as well, since the probability of generating a degenerate vertex in Monte Carlo sampled data is vanishingly small.

The right panels in figure 3 depict the corresponding Delaunay tessellation of the same data. Since the Voronoi and Delaunay tessellations are dual to each other, one way to construct the Delaunay tessellation is to start from the Voronoi diagram and join all pairs

of generator points whose Voronoi polygons share a common Voronoi edge. If the Voronoi tessellation is non-degenerate, each Voronoi vertex belongs to exactly three Voronoi edges, which in turn define a triangular polygon in the Delaunay tessellation (see figure 3); for this reason the Delaunay tessellation is often referred to as a *triangulation.* The described procedure[10] also manifestly pairs up all Voronoi edges with their corresponding dual edges in the Delaunay tessellation; if *each* such pair of dual edges from the two tessellations has a common point, i.e., the dual edges cross each other, the Delaunay triangulation is known as Pitteway triangulation [65, 66]. As we shall see explicitly below in figure 4, our datasets will generally *not* lead to Pitteway triangulations; so it will be important to keep in mind that some dual pairs of edges may be slightly offset and not intersect each other.

In what follows, our discussion will often switch back and forth between the two types of tessellations, so at this point it may be useful to build some intuition by recapping some of the relationships between the constituent objects of the Voronoi and Delaunay tessellations for a dataset consisting of N points:

- Each data point defines both a Voronoi polygon *and* a corresponding Delaunay vertex; the total number of Voronoi polygons or Delaunay vertices is thus N. The number of sides of a Voronoi polygon is equal to the number of Delaunay edges joining at the corresponding Delaunay vertex.

- Each Voronoi edge has a dual Delaunay edge; the total number of Voronoi edges is therefore equal to the total number of Delaunay edges and is on the order of, but slightly less than, $3N$ [42].

- Each Voronoi vertex defines a corresponding Delaunay triangle; the total number of such objects is on the order of, but slightly less than, $2N$ [42].

In summary, we have the following duality relations between the elements of the Voronoi and Delaunay tessellations:

$$\text{Voronoi cell} \longleftrightarrow \text{Delaunay vertex,} \tag{2.8a}$$

$$\text{Voronoi edge} \longleftrightarrow \text{Delaunay edge,} \tag{2.8b}$$

$$\text{Voronoi vertex} \longleftrightarrow \text{Delaunay triangle.} \tag{2.8c}$$

We can formalize these relations by introducing some notation. Let us use Latin indices to label elements from the Voronoi tessellation and Greek indices to label elements from the Delaunay triangulation. Then let $\{P_i\}$ be the set of generator points, $\{V_i\}$ be the set of Voronoi cells, $\{V_{ij}\}$ be the set of Voronoi edges and $\{V_{ijk}\}$ be the set of Voronoi vertices. Note that each Voronoi edge can be uniquely identified by the labels i and j of the pair of Voronoi cells which it separates, and similarly, for non-degenerate tessellations, a Voronoi vertex V_{ijk} is labelled by exactly three indices since it is the meeting point of the edges V_{ij}, V_{jk} and V_{ki}. Also let N_i be the number of edges (or equivalently, neighboring polygons) of the ith Voronoi cell V_i. With regards to the Delaunay triangulation, let $\{D_\alpha\}$ be the

---

[10]There are alternative methods to construct the Delaunay triangulation, e.g., using the property that the interiors of the circumcircles of Delaunay triangles are empty circles, i.e., contain no points from the dataset.

set of Delaunay triangles, $\{D_{\alpha\beta}\}$ be the set of Delaunay edges and $\{D_{\alpha_1\alpha_2\alpha_3...\alpha_{N_i}}\}$ be the set of Delaunay vertices. As before, each edge $D_{\alpha\beta}$ can be identified by the labels α and β of the Delaunay triangles which it separates, and each vertex can be identified by the labels $\{\alpha_1, \alpha_2, \ldots, \alpha_{N_i}\}$ of the N_i Delaunay triangles which are sharing it. Then the duality relations (2.8) can be written as

$$P_i \longleftrightarrow V_i \longleftrightarrow D_{\alpha_1\alpha_2\alpha_3...\alpha_{N_i}}, \tag{2.9a}$$

$$V_{ij} \longleftrightarrow D_{\alpha\beta}, \tag{2.9b}$$

$$V_{ijk} \longleftrightarrow D_{\alpha}, \tag{2.9c}$$

$$N_i \longleftrightarrow N_{\alpha} = 3. \tag{2.9d}$$

Note the trade-off in complexity — in the Voronoi case, vertices are labelled with exactly 3 indices, but the number of polygon edges N_i varies, while in the Delaunay case, the number of polygon edges is always 3, but vertices are labelled with a varying number of indices.

2.3   Candidate boundary objects

Having described the Voronoi and Delaunay tessellations of the data, before proceeding to the next two stages of gradient computation and tagging, it is worth pausing for a moment to discuss which elements of the tessellation are best suited for describing a wombling boundary. In figure 3 the theoretical boundary at $x = 0.5$ was marked with a vertical blue dashed line, but this was done only to guide the eye, since in reality both the existence and location of the boundary will be a priori unknown. In practice, we need to identify individual elements of the tessellations located at (or close to) the boundary which could be targeted by the wombling analysis.

By including the Delaunay tessellation into our discussion, we obtain three new possibilities in addition to the approach of refs. [16, 17]. The four panels in figure 3 illustrate these four options:

- *Boundary Voronoi Cells (BVCs)*, shown in the upper left panel in figure 3. When working with the Voronoi tessellation, this is the most natural and perhaps only option. A Boundary Voronoi Cell was defined as any Voronoi cell which is crossed by the theoretical boundary [16, 17]. In the upper left panel of figure 3 and in the left panel of figure 4, the BVCs are shaded in yellow.

- *Boundary Delaunay Triangles (BDTs)*, shown in the upper right panel in figure 3. When working with the Delaunay triangulation, in analogy we can now define a Boundary Delaunay Triangle to be any Delaunay triangle which is crossed by the theoretical boundary. In the upper right panel of figure 3 and in the right panel of figure 4, the BDTs are shaded in yellow.

- *Boundary Delaunay Edges (BDEs)*, shown in the lower right panel in figure 3. Similarly, we can define a Boundary Delaunay Edge to be any Delaunay edge which is crossed by the theoretical boundary. In the lower right panel of figure 3 and in the two panels of figure 4, the BDEs are indicated with green lines.
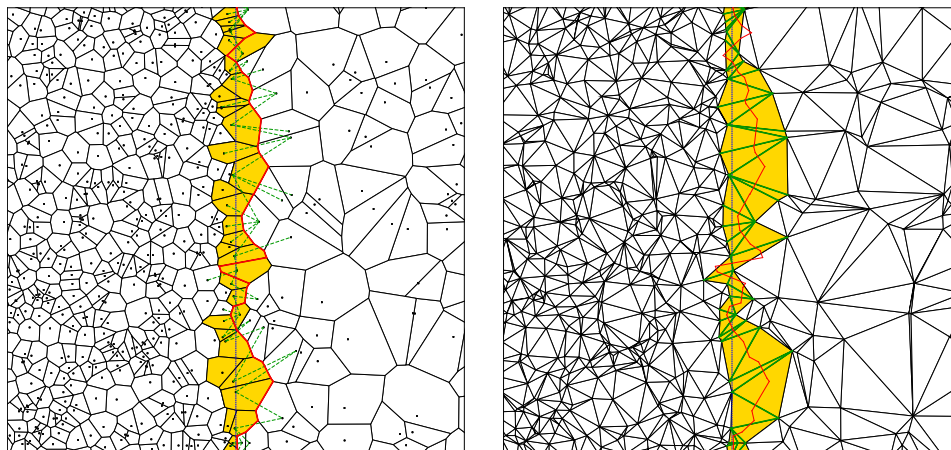
Figure 4. Left: the Voronoi tessellation of the data shown in figure 3 illustrating the relationships between Boundary Voronoi Cells (yellow-shaded), Boundary Voronoi Edges (red solid lines) and Boundary Delaunay Edges (green dashed lines). Right: Delaunay tessellation of the data shown in figure 3 illustrating the relationships between Boundary Delaunay Triangles (yellow-shaded), Boundary Voronoi Edges (red solid lines) and Boundary Delaunay Edges (green solid lines).

- *Boundary Voronoi Edges (BVEs)*, shown in the lower left panel in figure 3. Finally, using the duality relation (2.8b) we can define a Boundary Voronoi Edge to be any Voronoi edge which is dual to a Boundary Delaunay Edge. In the lower left panel of figure 3 and in the two panels of figure 4, the BVEs are indicated with red lines.

Note that the last three options all rely on the Delaunay triangulation and would not have been possible if we were only considering the Voronoi tessellation of the data.

Given the duality relations (2.8) between the Voronoi and Delaunay tessellations, the different categories of boundary objects defined above are related to each other. These relationships are exhibited in figure 4, where we simply superimpose some of the results from figure 3 in order to better see the existing correlations. Figures 3 and 4 demonstrate that all four definitions lead to a contiguous set of boundary objects strung along the theoretical boundary. Now the question becomes how to tag these boundary objects with a suitable algorithm using their geometric properties.

3 Estimation of local gradients from the tessellation

3.1 Gradient estimation from a Voronoi tessellation

As discussed in the Introduction, the Voronoi tessellation provides a natural estimate (1.5) for the *values* of the function f at the location of each generator point. In our case, since we are dealing with a two-dimensional dataset (2.1), eq. (1.5) reduces to

$$f_i \sim \frac{1}{a_i}, \tag{3.1}$$

where a_i is the area of the ith Voronoi cell V_i. The identification (3.1) is pictorially illustrated in figure 5 for the point data examples from figure 1. Unfortunately, the area
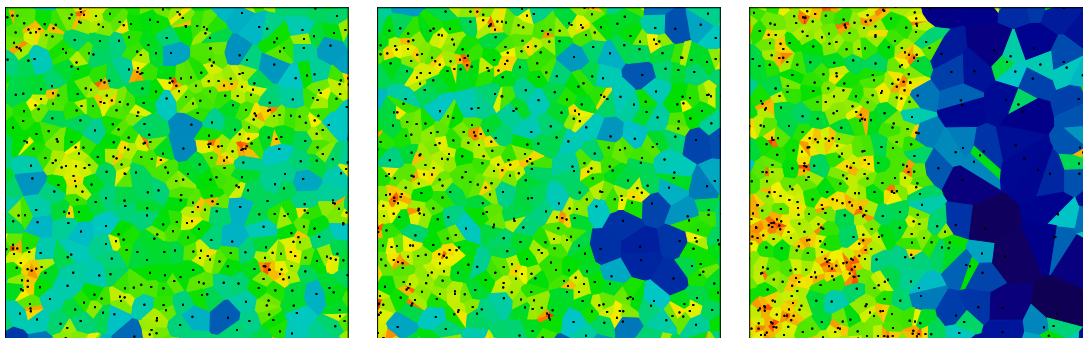
Figure 5. Pictorial illustration of the identification (3.1) for the case of the three point datasets shown in figure 1. The Voronoi cells are color-coded by their inverse areas with warm (cold) colors indicating large (small) values of f.

a_i by itself does not tell us anything about the *gradient* of the function $f(x, y)$ — for this purpose, we need to compare a_i to the areas of the surrounding Voronoi cells. Let $\mathcal{N}_i = \{j_1, j_2, \ldots, j_{N_i}\}$ be the set of indices labelling the neighboring Voronoi cells, i.e., the Voronoi cells sharing an edge with V_i. By taking the neighboring cells one at a time, $j \in \mathcal{N}_i$, one can compute directional derivatives $(\nabla_{\hat{n}_{ij}} f)_i$ in the direction of the jth neighboring cell, i.e., along the unit vector

$$\hat{n}_{ij} = \frac{1}{\sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}} (x_j - x_i, y_j - y_i),$$

as [16]

$$(\nabla_{\hat{n}_{ij}} f)_i = (a_i a_j)^{\frac{3}{4}} \frac{f_j - f_i}{\sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}}, \tag{3.2}$$

where the prefactor of $(a_i a_j)^{\frac{3}{4}}$ was included to make the directional derivative dimensionless. Since each Voronoi cell V_i has a varying number of edges $N_i \equiv |\mathcal{N}_i|$, there will be a different number of directional derivatives available at each point i, but they can all be fitted to the expected distribution from the true gradient, thus producing an estimate $\vec{G}_i$ of the gradient vector at the ith generator point P_i [16]. Another variable explored in ref. [16] was the relative standard deviation $\bar{\sigma}_i$ of the areas of the neighboring cells,

$$\bar{\sigma}_i \equiv \frac{\sigma_i}{\bar{a}_i} \equiv \frac{1}{\bar{a}_i} \sqrt{\sum_{j \in \mathcal{N}_i} \frac{(a_j - \bar{a}_i)^2}{N_i - 1}}, \tag{3.3}$$

where

$$\bar{a}_i \equiv \frac{1}{N_i} \sum_{j \in \mathcal{N}_i} a_j$$

is the mean area of the neighbors of the ith Voronoi cell. As demonstrated in refs. [16, 22, 56], among the different possibilities, the relative standard deviation (3.3) performed rather well in tagging the BVCs. Of course, those studies were utilizing only the Voronoi tessellation, which is not ideal for computing gradients. The Delaunay triangulation, on

the other hand, provides a more natural framework for the gradient estimation,[11] as will be discussed in the following subsection.

3.2 Triangulation wombling from a Delaunay tessellation

The Delaunay triangulation leads to a natural method for computing the gradient known as "triangulation wombling" [34, 39]. The starting point is the observation that for each Delaunay triangle D_α, the functional values at its three vertices are known from (3.1). Three points are enough to fit a plane, whose slope will provide an estimate of the gradient vector $\vec{G}_\alpha$ to be associated with the Delaunay triangle D_α. Recall the duality relation (2.9c) which maps the triangle D_α to its three vertices which carry indices i, j and k in the Voronoi tessellation. We can then parametrize the plane defined by D_α as

$$f = G_{\alpha x}\, x + G_{\alpha y}\, y + C, \tag{3.4}$$

where C is some constant. Applying this relation at each vertex, we obtain three independent equations

$$f_i = G_{\alpha x}\, x_i + G_{\alpha y}\, y_i + C, \tag{3.5a}$$

$$f_j = G_{\alpha x}\, x_j + G_{\alpha y}\, y_j + C, \tag{3.5b}$$

$$f_k = G_{\alpha x}\, x_k + G_{\alpha y}\, y_k + C, \tag{3.5c}$$

which can be solved for the gradient $\vec{G}_\alpha$ and the constant C as [34, 39]

$$\begin{pmatrix} G_{\alpha x} \\ G_{\alpha y} \\ C \end{pmatrix} = \begin{pmatrix} x_i & y_i & 1 \\ x_j & y_j & 1 \\ x_k & y_k & 1 \end{pmatrix}^{-1} \begin{pmatrix} f_i \\ f_j \\ f_k \end{pmatrix}. \tag{3.6}$$

From here, a wombling analysis would typically focus on the magnitude of the gradient

$$G_\alpha = \sqrt{G_{\alpha x}^2 + G_{\alpha y}^2} \tag{3.7}$$

and proceed to select Delaunay triangles D_α with relatively large values of G_α, typically in the top 10th percentile of all cells, as candidates for boundary elements.

However, the straightforward application of this procedure leads to a problem which is illustrated in figure 6 for the case of $\rho = 5$. In the left panel we show the distributions of gradient magnitudes G_α as calculated by the triangulation wombling method just described. We see that the distribution is a very steeply falling function with a long tail — most Delaunay cells have relatively small gradients and only a small fraction populates the large G_α tail. Now, if the cells on the tail were predominantly BDTs, the method would have succeeded and there would be no problem, but unfortunately, that is not the case. In the middle panel of figure 6 we show a scatter plot of the calculated gradient magnitudes G_α

---

[11]A purist might say that, since the two tessellations are dual to each other, strictly speaking there is nothing more to be gained from the Delaunay tessellation that could not have already be obtained from the Voronoi tessellation. While this may be technically correct, we found the Delaunay tessellation useful in hinting at some new techniques and ideas as discussed below.
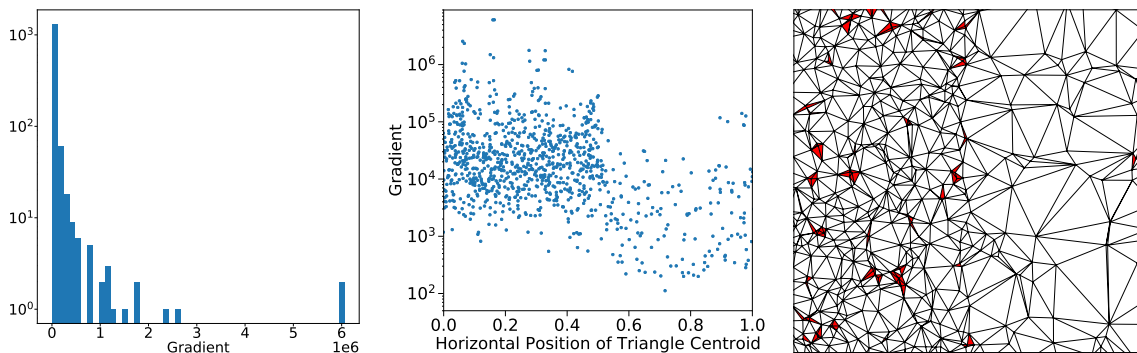
Figure 6. Left panel: distribution of the magnitudes G_α of the gradient vectors $\vec{G}_\alpha$ computed in the process of triangulation wombling. Middle panel: a scatter plot of G_α versus the horizontal position of the centroid of the corresponding Delaunay triangle D_α. Right panel: the Delaunay triangles (red-shaded) within the top 10th percentile as ranked by gradient magnitudes.

versus the horizontal position of the corresponding Delaunay triangle D_α as given by its centroid. We notice that the gradients computed for cells in the dense region ($x < 0.5$) tend to be much larger than the gradients in the sparse region ($x > 0.5$) (this is to be expected, since statistical fluctuations scale as $\sqrt{f}$). As a result, the cells with large gradients will tend to be more or less uniformly distributed in the dense region, with no relation to the boundary at $x = 0.5$. This is confirmed in the right panel of figure 6, where we identify with red shading the Delaunay triangles whose gradients are in the top 10th percentile of gradient magnitudes. As anticipated from the result in the middle panel, the red-shaded Delaunay triangles are located almost entirely in the dense region and there is no apparent clustering near the boundary. This means that in order to properly tag the BDTs, we must first pre-process the computed gradients $\vec{G}_\alpha$ in order to mitigate the effect of the statistical noise.

3.3 Denoising

In this subsection we outline three different procedures for denoising the computed gradient vectors $\vec{G}_\alpha$. As we shall demonstrate, each of them has the desired effect, and the optimal approach will be some combination of the three, although finding out the exact proportions is beyond the scope of this paper.

3.3.1 Rescaling of the naive gradients

The first approach is to rescale each calculated gradient magnitude G_α as

$$G_\alpha \to \tilde{G}_\alpha \equiv G_\alpha \sqrt{a_i a_j a_k}, \tag{3.8}$$

where a_i, a_j and a_k are the areas of the Voronoi cells centered on the three vertices of the Delaunay triangle D_α (recall the duality relation (2.9c)). The idea behind the rescaling (3.8) is to render the rescaled gradient $\tilde{G}_\alpha$ "dimensionless" with respect to the cell areas a_i, and therefore insensitive to the overall density of points. Figure 7 demonstrates that the rescaling (3.8) does have the desired effect. In the left panel, the distribution of rescaled gradient magnitudes $\tilde{G}_\alpha$ still follows the same trend as in figure 6, but the numerical values
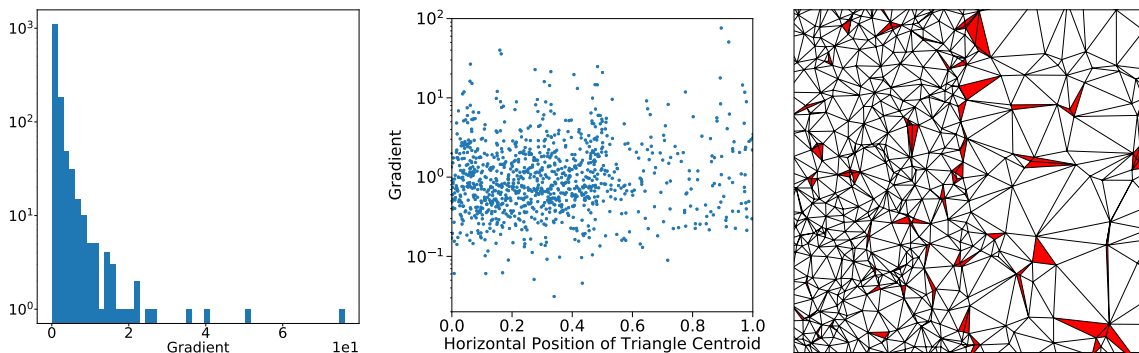
Figure 7. The same as figure 6, but after rescaling the calculated gradients as in eq. (3.8).

have been reduced by several orders of magnitude. More importantly, the middle panel of figure 7 confirms that the rescaled gradient magnitudes are now uniformly consistent across the two regions, with a cluster of points with relatively large $\tilde{G}_\alpha$ beginning to emerge near the theoretical boundary. The right panel of figure 7 shows the updated plot of the Delaunay triangles falling in the top 10th percentile of $\tilde{G}_\alpha$. We observe that, while the Delaunay cells tagged as BDTs are still scattered throughout the field of view, a significant fraction of them appears at the location of the boundary, which further validates the rescaling procedure (3.8). That is why from now on, unless specified otherwise, we shall always work with gradient vectors which have been rescaled as in (3.8).

3.3.2 Lloyd steps uniformization (LSU)

Another method for removing unwanted noise fluctuations in the data was explored in [16] and involved the so called Voronoi relaxation of the data by means of Lloyd's algorithm [67]. A closer inspection of the Voronoi tessellations depicted in figures 3 and 4 reveals that after the tessellation is constructed, the generator points can be found pretty much anywhere within the Voronoi polygon — near the center of the cell, close to an edge, or somewhere in between. The idea of the Voronoi relaxation is to make the whole Voronoi structure more uniform by performing several steps (or iterations) of Lloyd's algorithm, where at each iteration, the generator point is moved to the centroid of the corresponding Voronoi cell and the tessellation is redone.

The effect of performing such Lloyd step uniformization (LSU) on our data is shown in figures 8 and 9. Each figure has 8 panels, depicting the Delaunay tessellation[12] of the data after a certain number of Lloyd iterations, starting with 0 (no Lloyd steps) in the upper left panels and going up to 20 Lloyd steps in the lower right panels. In addition, in figure 8 each Delaunay triangle D_α is color-coded by the rescaled magnitude $\tilde{G}_\alpha$ of the respective gradient (the gradients are recalculated after each step). Note that the color bar extends only up to 60% of the largest gradient magnitude found in the data;[13] any cell with a gradient magnitude above that threshold is colored yellow, essentially creating

---

[12]For similar plots illustrating the effect of LSU on the Voronoi tessellation, see [16].

[13]For example, the left panel in figure 7 reveals that the largest $\tilde{G}_\alpha$ value found in the data is around 75, while the color bar in the upper left panel of figure 8 only goes up to $75 \times 60\% = 45$.
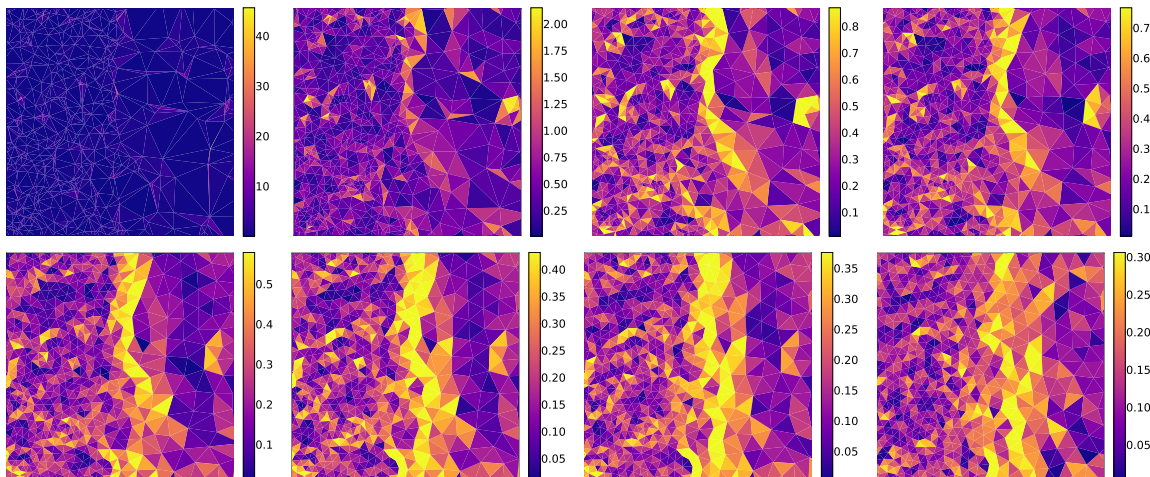
Figure 8. Delaunay triangulations with individual cells color-coded by their rescaled gradient magnitudes $\tilde{G}_\alpha$, after a certain number of Lloyd steps (from top left to bottom right): 0, 1, 2, 3, 5, 7, 10, 20. The color bar extends up to 60% of the largest $\tilde{G}_\alpha$ value found in the data; any cell with a gradient magnitude above that threshold is colored yellow, essentially creating an overflow color bin. This was done in order to minimize the effect of outliers and better visualize the bulk of the cells with typical values.
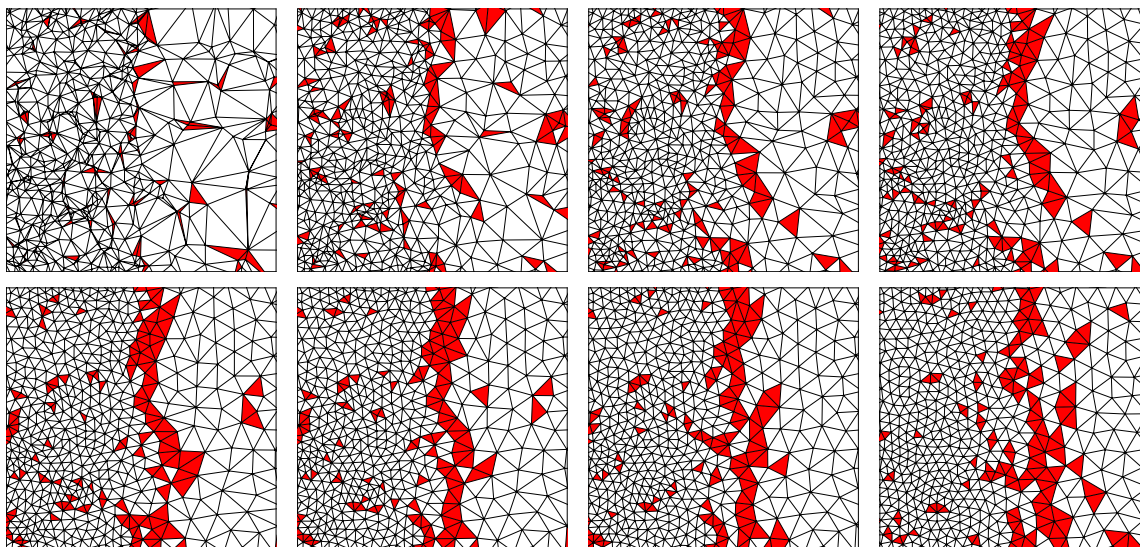
Figure 9. The same as the right panel in figure 7, but showing the effect of applying the Lloyd algorithm with varying number of steps as in figure 8.

an overflow color bin. This was done in order to minimize the effect of outliers and better visualize the bulk of the cells with the more typical values. Figure 9, on the other hand, marks the Delaunay triangles falling in the top 10th percentile of $\tilde{G}_\alpha$ values, just like the plot in the right panel of figure 7.

There are several lessons that can be drawn from figures 8 and 9. First, as expected, the Lloyd relaxation causes the Delaunay triangles to become more regularly shaped. For
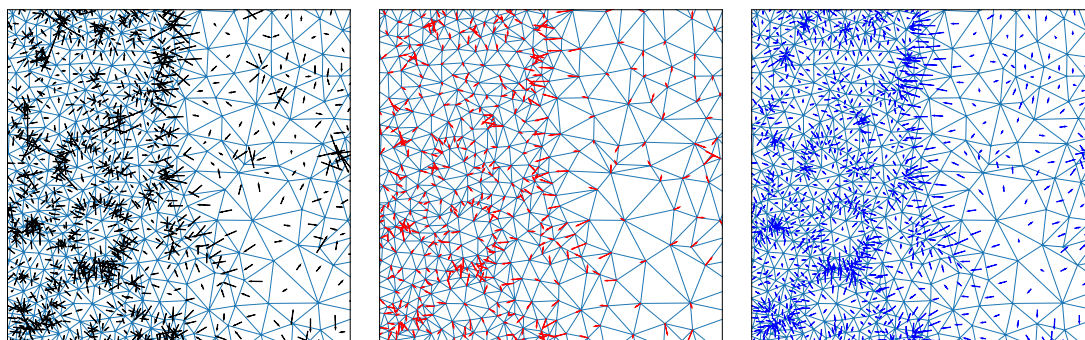
Figure 10. Left panel: rescaled gradient vectors $\tilde{\vec{G}}_\alpha$ computed via triangulation wombling, after applying one Lloyd iteration to the original data. Middle panel: the result $\langle \vec{G} \rangle_i$ from averaging the gradient vectors $\tilde{\vec{G}}_\alpha$ shown in the left panel at each data point P_i according to eq. (3.9). Right panel: the result $\langle \vec{G} \rangle_\alpha$ from further averaging the Voronoi-averaged gradient vectors $\langle \vec{G} \rangle_i$ shown in the middle panel at each Delaunay triangle according to eq. (3.10).

example, notice that a large fraction of the triangles in the original Delaunay tessellation are obtuse (see the right panels in figures 3–7). However, within a few Lloyd steps, the fraction of obtuse triangles drops significantly and obtuse triangles are rather rare[14] in the plots in the lower rows of figures 8 and 9. More importantly, as shown in figures 8, the LSU procedure also tends to wash out the noisy fluctuations in the calculated rescaled gradient magnitudes $\tilde{G}_\alpha$ within the bulk regions away from the boundary (note the decreasing range of $\tilde{G}_\alpha$ on the color bars). This further sharpens the contrast between the Delaunay triangles situated near the boundary versus those in the bulk. In particular, notice the gradual emergence of the boundary, which becomes quite pronounced and unmistakable after 5-7 Lloyd steps. However, the figures also show that the number of Lloyd steps should be chosen with care — applying too few may not optimally showcase the boundary, while applying too many may cause the boundary to start disintegrating, as evidenced in the lower right panels after 20 iterations.

3.3.3 Local averaging of gradient vectors

A third approach for smoothing out the local statistical fluctuations in the data is to perform some type of averaging procedure over a region extending beyond the individual Voronoi or Delaunay cells and their immediate neighbors. For example, ref. [16] considered extending the calculation of $\bar{\sigma}_i$ in (3.3) over several tiers of Voronoi neighbors (up to 5) and showed that this indeed produces the desired effect of reducing the fluctuations and sharpening the boundary identified by means of $\bar{\sigma}_i$.

In our case here, we are dealing with the Delaunay tessellation instead, where the procedures of triangulation wombling (3.6) and rescaling (3.8) allow us to directly compute the rescaled gradient vector $\tilde{\vec{G}}_\alpha$ associated with each Delaunay triangle D_α. The result (after one Lloyd iteration) is shown in the left panel of figure 10, where we plot each vector $\tilde{\vec{G}}_\alpha$ at the centroid of the corresponding Delaunay triangle. This plot is the vector field

---

[14]Furthermore, the largest angle of any remaining obtuse triangle is typically not too far above $90°$.
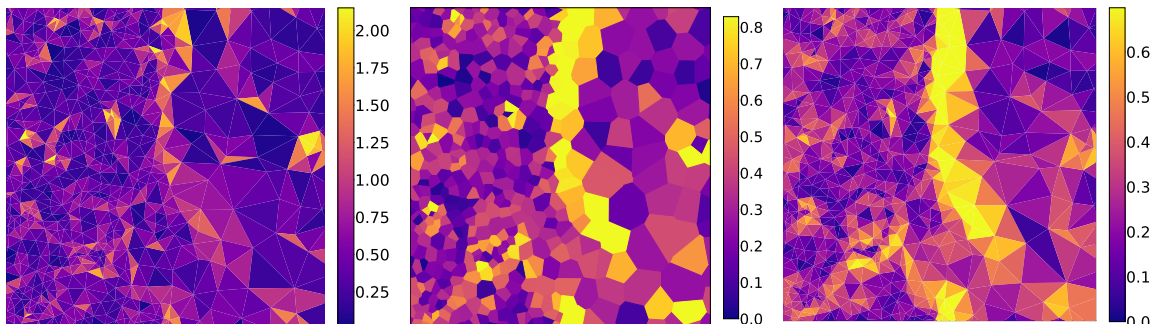
Figure 11. Delaunay (left and right panels) or Voronoi (middle panel) triangulations of the data after applying one Lloyd iteration, with individual cells color-coded by the magnitudes of their respective rescaled gradient vectors computed from triangulation wombling (left panel), Voronoi-averaged according to eq. (3.9) (middle panel), or Delaunay-averaged according to eq. (3.10) (right panel).

analogue of the color map shown in the second panel of the upper row of figure 8, which for ease of comparison we reproduce again in the left panel of figure 11. The difference is that the color map plots in figures 8 and 11 identify cells only by the magnitude $\tilde{G}_\alpha$ of the gradient, while the vector field plots of figure 10 include the directional information as well, which is useful in visualizing the spatial patterns and correlations of the gradient vectors.

Now, given the rescaled gradient vectors $\vec{\tilde{G}}_\alpha$ shown in the left panel of figure 10, there are two ways to perform local averaging of these vectors, depending on whether we want to associate the result from the averaging with a Voronoi cell (i.e., a Delaunay vertex) or with a Delaunay triangle:

- *Voronoi cell averaging.* Recall that according to the duality relation (2.9a), each Voronoi cell V_i corresponds to a Delaunay vertex $D_{\alpha_1 \alpha_2 \alpha_3 \ldots \alpha_{N_i}}$, which is the common vertex of N_i Delaunay triangles $D_{\alpha_1}, D_{\alpha_2}, D_{\alpha_3}, \ldots, D_{\alpha_{N_i}}$. Thus we can simply define the average gradient $\langle \vec{G} \rangle_i$ at any Voronoi cell V_i to be[15]

$$\langle \vec{G} \rangle_i = \frac{1}{N_i} \sum_{k=1}^{N_i} \vec{\tilde{G}}_{\alpha_k}. \tag{3.9}$$

- *Delaunay cell averaging.* Once we have the averaged vectors (3.9) at our disposal, we can go back to each Delaunay triangle and further average the three vectors (3.9) associated with its three vertices.

$$\langle \vec{G} \rangle_\alpha = \frac{1}{3} \left(\langle \vec{G} \rangle_i + \langle \vec{G} \rangle_j + \langle \vec{G} \rangle_k \right), \tag{3.10}$$

where the indices i, j and k label the data points at the vertices of D_α (recall the duality relation (2.9c)).

The vector fields resulting from the averaging prescriptions in eqs. (3.9) and (3.10) are shown in the middle and right panels of figure 10, respectively. One can see that the

---

[15]As an alternative to a simple sum as in (3.9), one could assign a weight for each vector $\vec{\tilde{G}}_{\alpha_k}$, for example, the angular size of the D_{α_k} triangle as seen from the point P_i.

statistical fluctuations are indeed getting suppressed as a result of the averaging, and furthermore, the directions of the gradient vectors near the boundary are becoming better correlated with each successive iteration. This directional correlation will become important in the next two sections where we shall compare the properties of *neighboring* cells in the tessellation. For now, in order to demonstrate the benefits from the averaging procedures (3.9) and (3.10) for the purposes of boundary detection, it is sufficient to update the color maps from figure 8 using the magnitudes of the *averaged* gradients instead. This is done in the middle and right panels of figure 11, where the individual cells in the tessellation have been color-coded by the magnitudes of the Voronoi-averaged gradient (3.9) and the Delaunay-averaged gradient (3.10), respectively. By comparing figure 11 to figure 8, we see that the local averaging procedures produce comparable benefits to Voronoi relaxation, so that we can view the two procedures as alternatives to the other. More specifically, local averaging seems to be at least equivalent to (if not better than) running on the order of 5-7 Lloyd iterations, which seemed to be the optimal choice in figures 8 and 9. Of course, the two methods can also be applied simultaneously, so that their benefits can be optimally exploited. In our analysis of sections 7 and 8, unless specified otherwise, we shall choose to employ the Delaunay-averaged gradient vectors (3.10).

4 Tagging elements of the tessellation as boundary candidates

Armed with the various estimates of the local gradient vectors discussed in the previous section, we are now ready for the next step of the wombling algorithm, namely, the tagging of Voronoi or Delaunay cells as boundary candidates. The standard approach is to place a lower cut on the relevant variable (typically the magnitude) which measures the size of the gradient. This selection singles out a certain set of candidate boundary cells as shown in figure 9. The purpose of this section is to study how effective[16] this selection is and to suggest a potential improvement of the standard approach by utilizing the correlations between gradient vectors computed in *neighboring* cells. The idea will be to place a premium not just on cells whose own gradient $\vec{G}_\alpha$ has a large magnitude, but on cells where the neighboring gradients $\tilde{\vec{G}}_\beta$ have both a) large magnitudes and b) correlated directions with $\tilde{\vec{G}}_\alpha$. A convenient variable which captures the desired correlations between two vectors $\tilde{\vec{G}}_\alpha$ and $\tilde{\vec{G}}_\beta$ is their dot product, $\tilde{\vec{G}}_\alpha \cdot \tilde{\vec{G}}_\beta$ [16].

Figure 12 shows scatter plots of such dot products of neighboring vectors for the three types of gradients introduced in the previous section: $\tilde{\vec{G}}_\alpha \cdot \tilde{\vec{G}}_\beta$ (upper right panel), $\langle \vec{G} \rangle_i \cdot \langle \vec{G} \rangle_j$ (lower left panel) and $\langle \vec{G} \rangle_\alpha \cdot \langle \vec{G} \rangle_\beta$ (lower right panel). In each case, the result is plotted versus the horizontal position $(x_i + x_j)/2$ of the midpoint of the respective[17] Delaunay edge

---

[16]Selection efficiency is typically illustrated with ROC curves, where one varies the cut on the selection variable and plots the fraction of signal versus the fraction of background surviving the cut. This was also the procedure used in refs. [16, 22]. Here, however, we prefer to simply show scatter plots of the tagging variable versus the distance to the boundary. In this way, we avoid the need to define what exactly is meant by a "boundary" object versus a "bulk" object.

[17]In the case of $\tilde{\vec{G}}_\alpha \cdot \tilde{\vec{G}}_\beta$ and $\langle \vec{G} \rangle_\alpha \cdot \langle \vec{G} \rangle_\beta$, the relevant Delaunay edge is simply $D_{\alpha\beta}$, i.e., the edge separating the Delaunay triangles D_α and D_β, while in the case of $\langle \vec{G} \rangle_i \cdot \langle \vec{G} \rangle_j$ the relevant Delaunay edge is the one dual to the Voronoi edge V_{ij}, see eq. (2.9b).
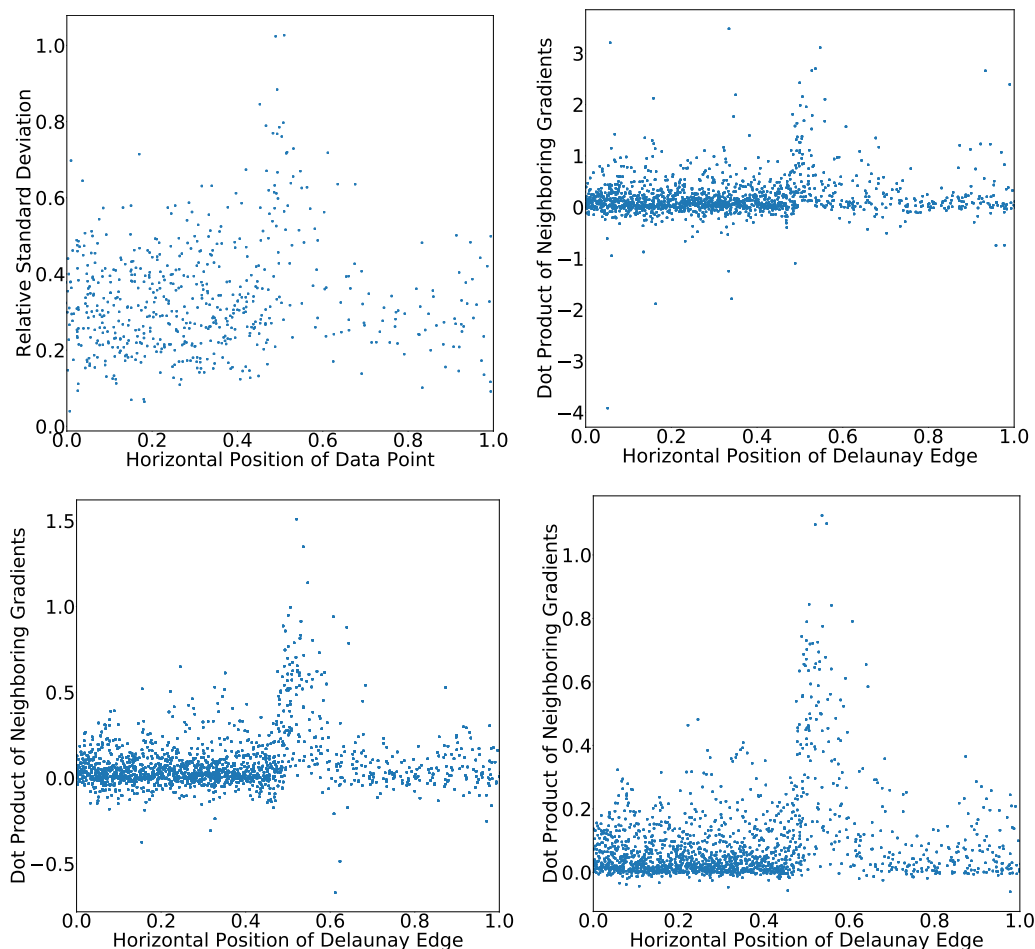
Figure 12. Scatter plots of different tagging variables discussed in the text versus the horizontal position of the corresponding element in the tessellation. The upper left panel shows the relative standard deviation $\bar{\sigma}_i$ defined in (3.3) versus the horizontal position x_i of the generator point P_i. The other three panels show dot products of different versions of neighboring gradient vectors versus the horizontal position $(x_i + x_j)/2$ of the midpoint of the respective Delaunay edge $D_{\alpha\beta}$: the rescaled gradients $\tilde{\vec{G}}_\alpha \cdot \tilde{\vec{G}}_\beta$ (upper right panel), the Voronoi-averaged gradients $\langle\vec{G}\rangle_i \cdot \langle\vec{G}\rangle_j$ (lower left panel) and the Delaunay-averaged gradients $\langle\vec{G}\rangle_\alpha \cdot \langle\vec{G}\rangle_\beta$ (lower right panel).

$D_{\alpha\beta}$. For comparison, in the upper left panel we show a scatter plot of the relative standard deviation $\bar{\sigma}_i$ defined in (3.3) versus the horizontal position x_i of the corresponding generator point P_i. As explained in section 3.1, the relative standard deviation $\bar{\sigma}_i$ is constructed from the Voronoi tessellation and was found to perform best among several other Voronoi-constructed alternatives [16]. The upper left panel in figure 12 confirms that the highest values of $\bar{\sigma}_i$ are indeed found for cells near the boundary; in fact, the top 16 highest $\bar{\sigma}_i$ values belong to such cells. At the same time, we also observe a significant variation in the $\bar{\sigma}_i$ values for cells in the bulk; for $\bar{\sigma}_i < 0.7$ this starts introducing a certain number of false positives.

The remaining three panels of figure 12 demonstrate that the corresponding dot products of gradients computed from the Delaunay tessellation are also efficient in identifying
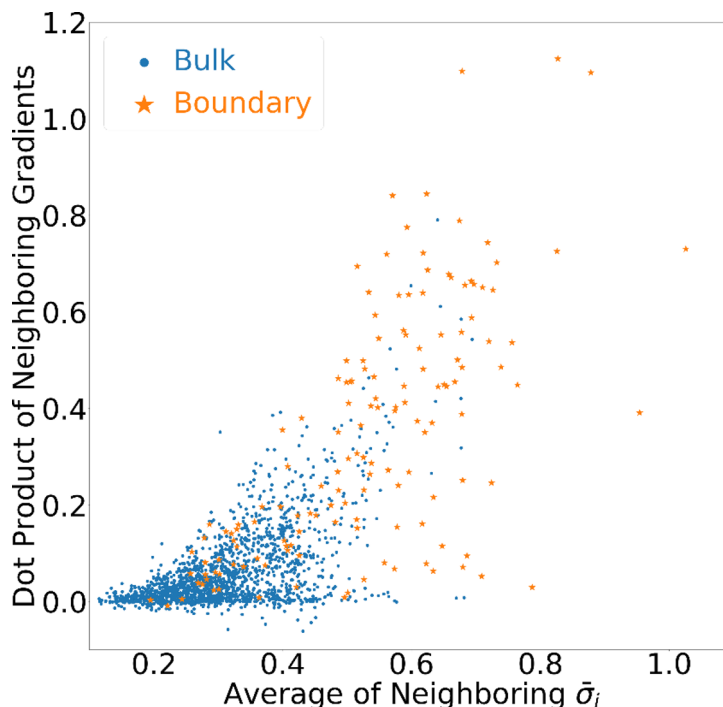
Figure 13. Scatter plot showcasing the correlation between two of the tagging variables from figure 12 — the relative standard deviation (x-axis) and the dot product $\langle \vec{G} \rangle_\alpha \cdot \langle \vec{G} \rangle_\beta$ of Delaunay-averaged gradients (y-axis). The orange stars correspond to Delaunay edges $D_{\alpha\beta}$ whose midpoint is located close to the boundary ($0.47 < x < 0.57$), while blue circles represent the remaining edges further away. Since each Delaunay edge $D_{\alpha\beta}$ is dual to two Voronoi points P_i and P_j, the quantity shown on the x-axis is the average $(\bar{\sigma}_i + \bar{\sigma}_i)/2$ of the respective relative standard deviations $\bar{\sigma}_i$ and $\bar{\sigma}_j$.

boundary objects (in this case, Delaunay edges). Among the three options illustrated in the plot, the dot product of the Delaunay-averaged gradients seems to perform the best — the top 47 highest dot products of neighboring $\langle \vec{G} \rangle_\alpha$ vectors belong to Delaunay edges near the boundary; and there is a well defined cluster of points with large y values in the boundary region $0.45 < x < 0.6$. Note the different y-axis range on these three plots — the variation of dot product values is largest for $\tilde{\vec{G}}_\alpha \cdot \tilde{\vec{G}}_\beta$ and smallest for $\langle \vec{G} \rangle_\alpha \cdot \langle \vec{G} \rangle_\beta$, further demonstrating the beneficial effects from the averaging procedures discussed in section 3.3.3.

Comparing the top left panel in figure 12 to the other three panels, we conclude that the gradient dot products, which take advantage of the correlations between neighboring gradient vectors in terms of both direction and magnitude, are able to identify the boundary better than $\bar{\sigma}_i$ and similar variables (figure 13 shows a direct comparison between the relative standard deviation and the dot product $\langle \vec{G} \rangle_\alpha \cdot \langle \vec{G} \rangle_\beta$ of Delaunay-averaged gradients). As a byproduct of this new method of tagging, we have also automatically built up a network of associations among the Delaunay cells in the triangulation, which is readily available for use in the next step (agglomeration), where one attempts to construct the actual boundary. This is the subject of the next section.
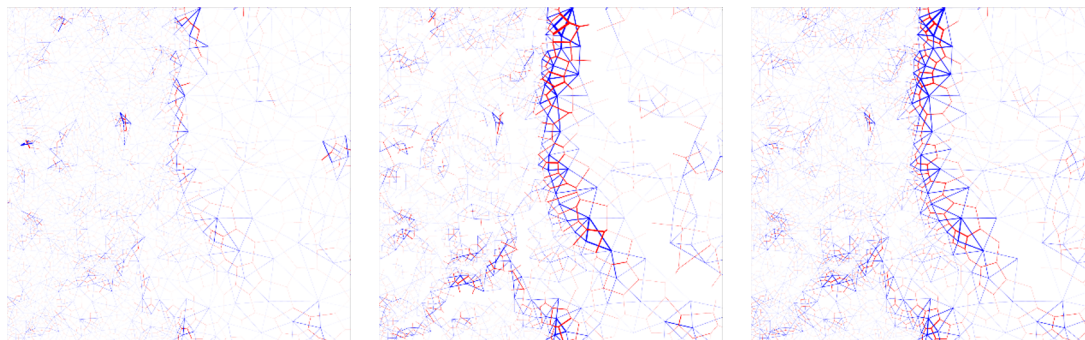
Figure 14. Weighted network representations of the Voronoi tessellation (red lines) and the Delaunay triangulation (blue lines). The line thickness is proportional to the weight, which is given by $\tilde{\vec{G}}_\alpha \cdot \tilde{\vec{G}}_\beta$ (left panel), $\langle \vec{G} \rangle_i \cdot \langle \vec{G} \rangle_j$ (middle panel) and $\langle \vec{G} \rangle_\alpha \cdot \langle \vec{G} \rangle_\beta$ (right panel). Edges with negative weights are not shown.

5 Agglomeration of tagged boundary elements

The previous step (the tagging of boundary elements) typically fails to result in a continuous boundary, especially in case of weak signals. Instead, the algorithm produces a collection of scattered "islands" of tagged cells, as seen in the right panels of figures 6 and 7 and in the top panels of figure 9. This necessitates the next step of agglomerating the individual tagged cells into subgraphs and evaluating whether the resulting pattern is consistent with a linear boundary [29]. The downside of this approach is that it does not take into further consideration the cells which have failed the tagging cut — those cells are simply ignored from this point on. Another potential drawback is that the tagging and agglomeration steps are done independently from each other, so that the existence of any spatial correlations among neighboring cells is not being used during the tagging. As already mentioned in the previous section, both of these problems are avoided when we use the dot products of neighboring gradients as tagging variables. As illustrated in the last three panels of figure 12, each dot product of gradients can be uniquely associated with a Delaunay edge $D_{\alpha\beta}$ or with its dual Voronoi edge V_{ij}, see eq. (2.9b). We can then treat the original Voronoi and Delaunay tessellations as *weighted networks*, where each edge is assigned a weight equal to the dot products of the corresponding two gradients. This weighted network representation is illustrated in figure 14, where we superimpose the Voronoi tessellation (red lines) and the Delaunay triangulation (blue lines). The weight of an edge is indicated by the line thickness — thicker lines imply higher weights and vice versa. The three panels show three different ways to compute the weights from the gradient dot products, depending on which set of gradient vectors from figure 10 we choose to use: $\tilde{\vec{G}}_\alpha \cdot \tilde{\vec{G}}_\beta$ (left panel), $\langle \vec{G} \rangle_i \cdot \langle \vec{G} \rangle_j$ (middle panel) or $\langle \vec{G} \rangle_\alpha \cdot \langle \vec{G} \rangle_\beta$ (right panel). As seen in figure 12, a certain fraction of dot products are negative; if that is the case, the corresponding edge is not plotted.

The three panels in figure 14 can be contrasted with the corresponding results in figure 11, where we used just the magnitudes of the individual gradient vectors, without any reference to their neighbors. The boundary seems to be better outlined in figure 14, particularly when we make use of the averaging procedures from section 3.3.3. We also

note the benefit of plotting the Voronoi and Delaunay tessellations simultaneously — the orientation of the edges with respect to the boundary is random, so whenever a given edge happens to be orthogonal to the boundary, its dual tends to be parallel to it, so taken together, they trace out the correct shape of the boundary.

Figure 14 also elucidates the results from figure 12, where we observed that while many edges situated close to the boundary enjoyed relatively large values of their gradient dot products, there was also a non-negligible fraction of edges near the boundary with rather low values of the gradient dot products. Figure 14 now reveals that these two populations are spatially correlated — note how the edges with large values of the gradient dot products are linked together, as are their counterparts. This confirms that using the gradient dot products as tagging variables automatically also takes care of the agglomeration.

Until now we have been following the standard steps of a wombling analysis. As already mentioned, the last remaining step is to evaluate the statistical significance of the observed pattern of tagged boundary candidates. Note that the last three figures illustrate tagging procedures for each of the four types of boundary candidate objects defined in section 2.3: Voronoi cells (middle panel in figure 11 and upper left panel in figure 12), Delaunay cells (left and right panels in figure 11), Delaunay edges (upper right and lower panels in figure 12 and all three panels in figure 14) and Voronoi edges (figure 14). Of course, since the Voronoi and Delaunay edges are dual to each other, any procedure which can tag one edge type can also be applied to tag the other.

In the next three sections we shall outline an alternative approach originally proposed in [40], which allows us to perform the tagging, agglomeration and statistical evaluation steps in one go. In order to introduce and illustrate the method, in the next section 6 we shall start with the case of a continuously defined function $f(x, y)$ and then proceed to analyze the case of point data in sections 7 and 8.

6 Finding wombling boundaries: analytical examples

In order to bypass the tagging and agglomeration steps, ref. [40] proposed to directly consider various curves C in the (x, y) plane, and to associate a "wombling measure" Γ with each one, so that true wombling boundaries can be identified by their large values of Γ. Since a wombling boundary is supposed to represent a zone of rapid change in the function f, it is natural to define the wombling measure in terms of the local gradient $\vec{\nabla} f$ suitably integrated along C. In particular, ref. [40] defined Γ to be the total gradient flux through C

$$\Gamma[C] \equiv \int_C \left(\vec{\nabla} f \cdot \hat{n}_C \right) d\ell, \tag{6.1}$$

where $\hat{n}_C$ is a unit vector normal to the curve C and $d\ell$ is the infinitesimal length along C. Additionally, ref. [40] also considered the average gradient flux

$$\bar{\Gamma}[C] \equiv \frac{\int_C \left(\vec{\nabla} f \cdot \hat{n}_C \right) d\ell}{\int_C d\ell}, \tag{6.2}$$

where using the total length of the curve as a normalization factor eliminates the unfair advantage of curves which happen to be too long. For this reason, in what follows we shall make use of (6.2) and not (6.1).[18]

Without any further constraints on the type of curves C that we are allowed to consider, this method would be rather impractical. To make further progress, two approaches are possible. The first one is the model-dependent route — if we specify exactly what kind of new physics model generates the wombling boundary, then C can be specified by only a handful of parameters (typically the masses of the new particles). Then the problem of maximizing the functional (6.2) over all possible curves C reduces to a simple global maximization problem in the parameter space describing C [23]. Here, however, we would like to stay as model-independent as possible, so we shall not assume any specific parametrization of the boundary. At the same time, we do not want to consider arbitrarily general curves C either.

An intermediate compromise approach is the following. Note that any curve C can locally be approximated by a straight line segment. Therefore, we can perform a scan of the (x, y) plane where at each point P we try a line segment (centered on P) of fixed length L and arbitrary angular orientation φ. Each point in the so-defined 4-dimensional parameter space (x, y, L, φ) corresponds to a well-defined line segment, for which the wombling measure (6.2) can be calculated. The regions in (x, y, L, φ) parameter space with large values for $\bar{\Gamma}[C]$ would then identify (segments of) the wombling boundary. Since this procedure involves optimization in a 4-dimensional parameter space, it will be difficult to illustrate here. This is why from now on we choose to focus on the (L, φ) subspace — one can think of this as first zooming in on an interesting region of the (x, y) plane and then testing for the presence of a linear wombling line segment.

Our reparametrization of the remaining two degrees of freedom describing the line segment is illustrated in figure 15. As before, we retain the unit square as our field of view. We then consider all possible straight lines crossing the unit square — each such line can be identified by the point where it enters the square and the point where it exits the square. We shall identify the locations of those two points by their respective coordinates p_{in} and p_{out} measured along the perimeter, as shown in the left panel of figure 15. Since the perimeter of a unit square is equal to 4, the parameter space $(p_{\text{in}}, p_{\text{out}})$ spans the 4×4 square shown in the right panel of figure 15 — any point within that 4×4 square can be uniquely associated with a straight line crossing the field of view in one of the two possible directions. For example, $(p_{\text{in}}, p_{\text{out}}) = (0, 2)$ describes a diagonal line traversed from the lower left corner to the upper right corner, while $(p_{\text{in}}, p_{\text{out}}) = (2, 0)$ describes the same

---

[18]An additional variation mentioned in [40] was to consider integrating the flux in absolute value, e.g.,

$$\Gamma[C] \equiv \int_C \left| \vec{\nabla} f \cdot \hat{n}_C \right| d\ell,$$

in order to avoid cases where a large positive flux over one section of the curve C is cancelled by a large negative flux over another section. However, in our case such cancellations are welcome since the noise fluctuations are random and we would like to allow them to cancel out each other as much as possible. We have confirmed numerically in our examples that the absolute value alternative leads to lower sensitivity.
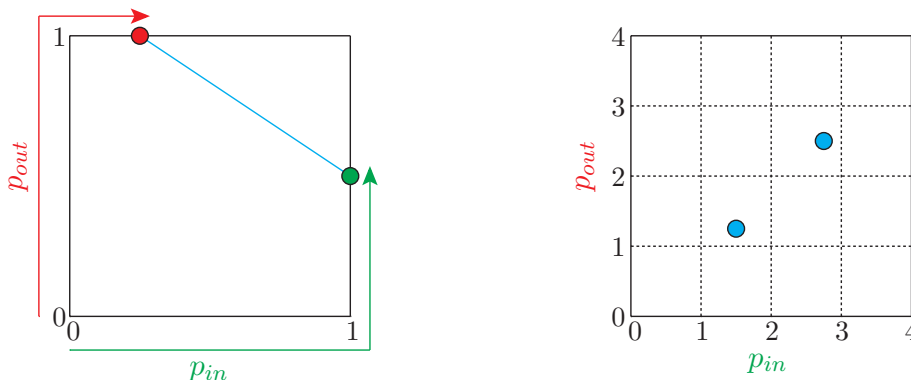
Figure 15. Our parametrization of all possible straight lines intersecting the unit square in terms of the coordinate p_{in} of the entry point and the coordinate p_{out} of the exit point, where p_{in} (p_{out}) is measured counterclockwise (clockwise) along the perimeter. The right panel shows the complete $(p_{\text{in}}, p_{\text{out}})$ parameter space, which is a square of side length 4. The blue dot represents the blue example line shown in the left panel and its partner (when the line is traversed in reverse).

diagonal line covered in reverse. In our previous examples, the true boundary was located at $x = 0.5$, and corresponds to either[19] $(p_{\text{in}}, p_{\text{out}}) = (0.5, 1.5)$ or $(p_{\text{in}}, p_{\text{out}}) = (2.5, 3.5)$.

In the remainder of this section and in the next section 7, our main goal will be to compute the wombling measure $\bar{\Gamma}[C]$ in the $(p_{\text{in}}, p_{\text{out}})$ parameter space and identify the relevant wombling boundary segment(s). First we shall illustrate this procedure with the example of a continuous function $f(x, y)$ before tackling the case of point datasets in the next section.

6.1 A straight line boundary

In this subsection we shall revisit the vertical straight line boundary example from the previous sections. However, we shall not use the original distribution (2.5), for two reasons: first, the discontinuity at $x = 0.5$ would generate an infinite gradient when computed analytically, and second, the distribution (2.5) corresponds to an idealized situation where the effects of particle widths and detector smearing are ignored. In any realistic experimental analysis the sharp step at $x = 0.5$ will be smeared and the boundary will be characterized by a large but finite gradient.

This is illustrated in figure 16, where the black dotted lines show (the x-dependence of) the original distribution (2.5) before any smearing, for $\rho = 5$ (left panel) and $\rho = 1.5$ (right panel). We then apply Gaussian smearing with $\sigma = 0.2$, resulting in the red histograms, which have the typical shapes expected in a realistic experimental analysis. In particular, notice how the gradient at the boundary is significantly reduced as a result of the smearing, making the task of finding the wombling boundary quite challenging. At this point we fit an analytical function to the so obtained smeared distributions. For the fit, we choose to

---

[19]In our conventions, reversing the direction of a given line $(p_{\text{in}}, p_{\text{out}})$ implies $(p_{\text{in}}, p_{\text{out}}) \to (4 - p_{\text{out}}, 4 - p_{\text{in}})$.
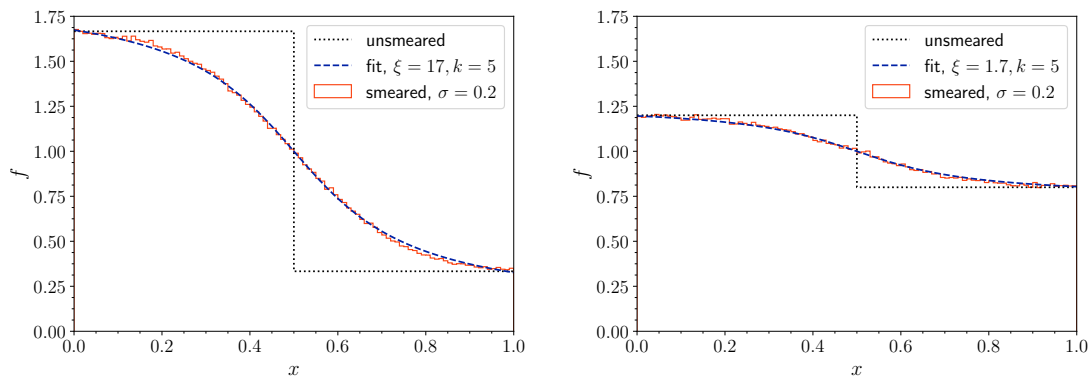
Figure 16. The black dotted line shows the original distribution (2.5) for $\rho = 5$ (left panel) and $\rho = 1.5$ (right panel) before any smearing. The red solid histogram is the result from applying Gaussian smearing with $\sigma = 0.2$. The blue dashed line is the result from fitting to the ansatz (6.4).

utilize the (unit-normalized) ArcTan sigmoid function

$$g_{\text{res}}(x; a, \xi, k) = 1 - \frac{2}{\pi} \frac{\xi - 1}{\xi + 1} \arctan\left(k(x - a)\right), \tag{6.3}$$

whose derivative is maximal (in absolute value) at $x = a$, the sharpness of the transition being controlled by the parameter k. The remaining parameter ξ is analogous to ρ in the sense that (compare to (2.4))

$$\xi = \frac{g(x = -\infty)}{g(x = \infty)}.$$

Since our field of view is limited to $x \in [0, 1]$ and the boundary is at $x = 0.5$, for the actual fit we choose the parametrization

$$f(x, y) = g_{\text{res}}(x; a = 0.5, \xi, k) \tag{6.4}$$

and then adjust ξ and k to match the smeared distributions shown by the red histograms. As seen in figure 16, the fit reproduces the effects of smearing rather well, so in the rest of this subsection we shall use (6.4) as our analytically defined distribution.

Using the respective fit (6.4) as our proxy, we can now compute the wombling measure $\bar{\Gamma}[C]$ in the $(p_{\text{in}}, p_{\text{out}})$ space of all lines intersecting our unit square. The result for $\rho = 5$ ($\rho = 1.5$) is shown in the left (right) panel of figure 17. We choose to plot the absolute value of $\bar{\Gamma}[C]$, since the sign of $\bar{\Gamma}[C]$ is determined by the direction in which we traverse the line, and does not have any bearing on whether the line is a wombling boundary or not. Figure 17 reveals that, as expected, there are two locations with maximal wombling measures: at $(p_{\text{in}}, p_{\text{out}}) = (0.5, 1.5)$ and at $(p_{\text{in}}, p_{\text{out}}) = (2.5, 3.5)$. Both of those correspond to the same vertical line at $x = 0.5$ which was the true boundary. This demonstrates that the method is indeed able to find the correct boundary. The significance of these findings, however, is sensitive to the amount of signal present — in the left panel, where $\rho = 5$, the two winning answers are very clearly identified, while in the right panel (using the same color scale) they appear to be less noticeable, which is a hint that the effect might be in danger of being washed out once we include the statistical fluctuations present in the point data — this issue will be investigated in detail in section 7 below.
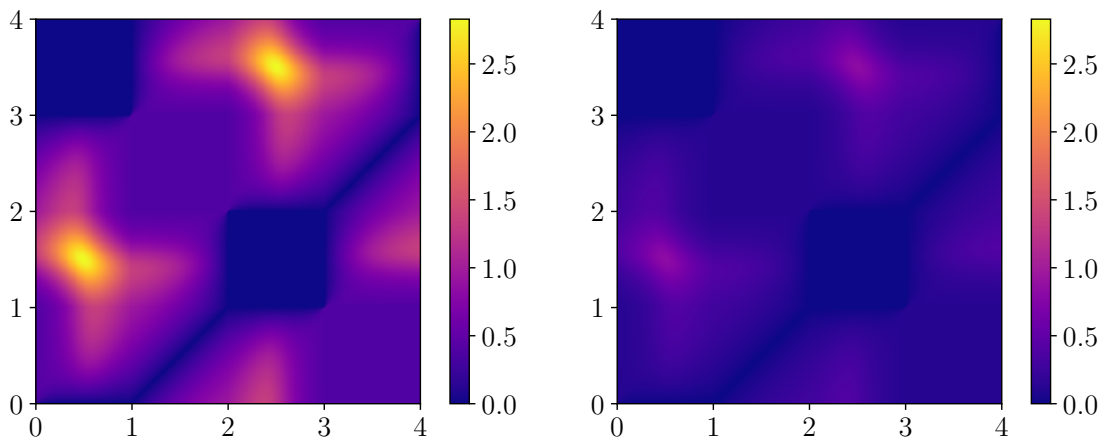
Figure 17. The result from the linear wombling procedure described in the text on a data set with $\rho = 5$ (left) and $\rho = 1.5$ (right). The heat map is color coded according to the value of $|\bar{\Gamma}[C]|$ and exhibits two degenerate maxima since each line is traversed twice (once in each direction).

6.2 A circular boundary line

The example in the previous section 6.1 might appear somewhat contrived since the true boundary was a straight line and at the same time, we also used straight lines in computing the wombling measure $\bar{\Gamma}[C]$. Since the shapes of the lines match, it was inevitable to find a unique best match, as shown in figure 17. To be fair, we shall now consider a less trivial example where the true boundary has a different shape from the line segments which we use to test for the presence of a wombling boundary. In particular, we shall revisit the case of a circular boundary introduced in section 2.1, where the probability distribution was given by (2.7). Once again, we shall not rely directly on (2.7), but in order to account for the detector resolution, we shall sample the events according to

$$f(r, \varphi) \sim r g_{\text{res}}(r; a = 0.25, \xi = 2, k = 30), \tag{6.5}$$

where r and φ are the polar coordinates in the plane, measured from an origin at the center of the circle, and the smearing function g_{res} was already defined in (6.3). The resulting probability distribution is plotted in the left panel of figure 18. Note that the densities on the two sides of the circular boundary differ by no more than a factor of 2, so in this sense this example is analogous to $\rho \sim 2$ in our usual notation. The corresponding heat map of $|\bar{\Gamma}[C]|$ in the $(p_{\text{in}}, p_{\text{out}})$ parameter space is shown in the right panel of figure 18. The most striking difference from the previous results in figure 17 is that now we find not just a single wombling boundary candidate, but a whole class of wombling boundaries, identified by the two[20] bright yellow stripes running diagonally across the plot. A careful inspection of the right panel in figure 18 reveals that each of the identified wombling boundary candidates is tangential to the circular boundary, which suggests that the dominant contribution to the

---

[20]We obtain two stripes because of the double counting $(p_{\text{in}}, p_{\text{out}}) \longleftrightarrow (4 - p_{\text{out}}, 4 - p_{\text{in}})$ due to the possibility to traverse a line segment in each of the two opposite directions, as shown in the right panel of figure 15.
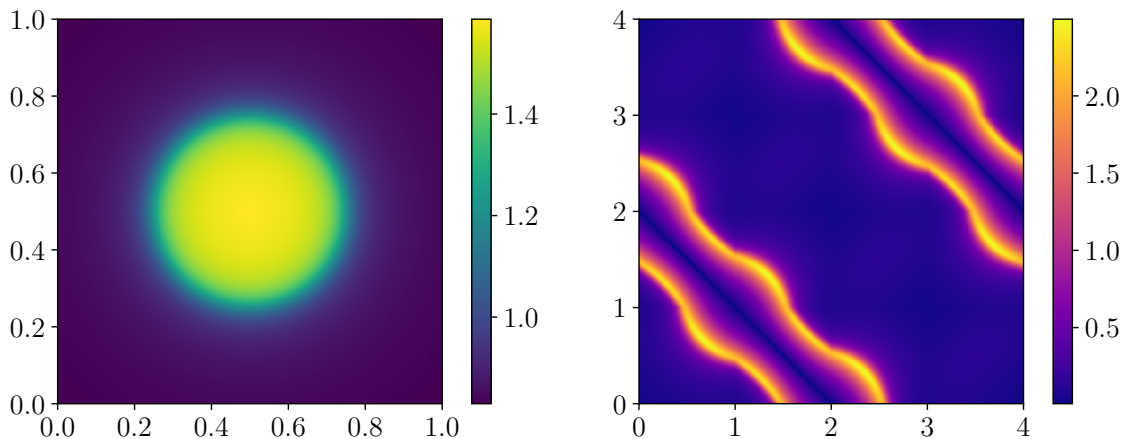
Figure 18. Color maps of the probability distribution (6.5) used for the circular boundary example (left panel) and of the corresponding $|\bar{\Gamma}[C]|$ in the $(p_{\mathrm{in}}, p_{\mathrm{out}})$ parameter space (right panel).

integral comes from the region in the vicinity of the circular boundary, where the gradient is largest. The slight difference in the brightness along the stripes can then be attributed to the different orientation of the lines, which leads to differences in the (sub-dominant) flux contributions in the regions away from the true boundary.

7 Finding wombling boundaries: point data examples

Having illustrated the basic idea of refs. [40, 63] with the continuous examples from the previous section, we shall now apply it to point data. Following the outline of section 6, we shall first consider the case of a straight line boundary in section 7.1 and then the case of a circular boundary in section 7.2.

7.1 A straight line boundary

7.1.1 An example with $\rho = 5$

Our first straight-line boundary example will be the same point data example which we have been using so far throughout the paper to illustrate the various methods and techniques of a wombling analysis, see figures 3, 4, 6, 7, 8, 9, 10, 11, 12 and 14. (As a reminder, we used $N = 500$ points generated according to the distribution (2.5) with $\rho = 5$.) In particular, we shall repeat the procedure from section 6 and compute a wombling measure $\bar{\Gamma}[C]$ for each possible straight line C crossing the field of view. However, since we are now dealing with discretely sampled point data instead of a continuous function $f(x, y)$, we need to adapt the definition (6.2) as follows

$$\bar{\Gamma}[C] \equiv \frac{\sum_\alpha \left(\langle \vec{G} \rangle_\alpha \cdot \hat{n}_C \right) \Delta \ell_\alpha}{\sum_\alpha \Delta \ell_\alpha},$$
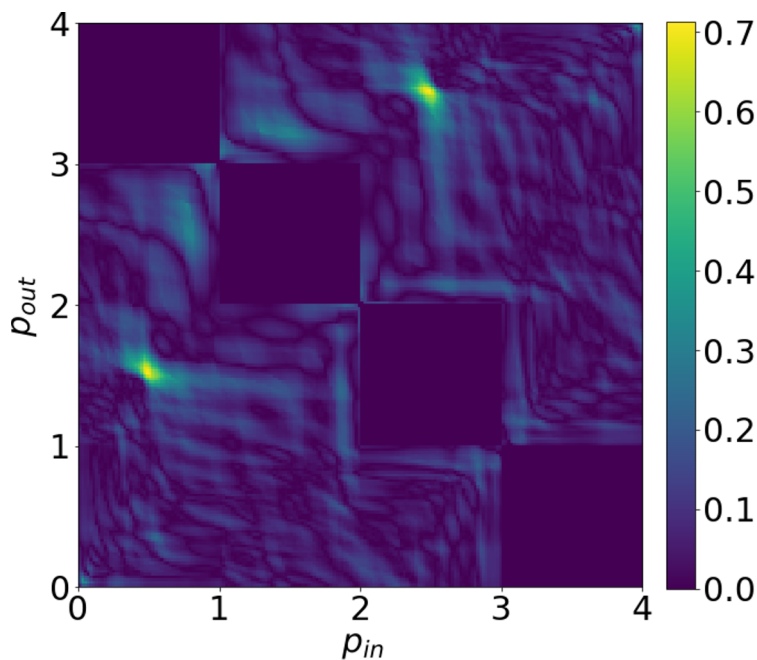
(7.1)

Figure 19. Results from applying the wombling procedure of section [6] to the illustrative point data example used in the previous sections ($N = 500$ points generated from (2.5) with $\rho = 5$). For concreteness, at the pre-processing stage we applied one Lloyd iteration and then used the Delaunay-averaged gradient vectors which were defined in (3.10). In constructing the heat map, we sampled the (p_{in}, p_{out}) parameter space on a 200 by 200 grid with step size 0.02 and then computed $\bar{\Gamma}[C]$ from eq. (7.1). The four dark squares in the plot correspond to line segments which overlap with one of the edges of the unit square which is our field of view — those segments were assigned $\bar{\Gamma}[C] = 0$ by default and excluded from further consideration. For reference, the true boundary is located at $(p_{in}, p_{out}) = (0.5, 1.5)$, or equivalently, at $(p_{in}, p_{out}) = (2.5, 3.5)$.

where each sum runs over all Delaunay triangles D_α which are crossed by the straight line C and, as before, $\langle \vec{G} \rangle_\alpha$ is the Delaunay-averaged[21] gradient vector (3.10) associated with D_α. Since the Delaunay tessellation gives complete coverage of the field of view, the line C necessarily gets fragmented into individual line segments of length $\Delta \ell_\alpha$, defined so that each segment is contained within a single Delaunay cell D_α. Finally, $\hat{n}_C$ is a unit vector orthogonal to the straight line C and therefore, to each individual line segment $\Delta \ell_\alpha$ as well, so that an index α on it is unnecessary.

Figure 19 shows a heat map of the wombling measure $\bar{\Gamma}[C]$ computed from eq. (7.1) throughout the parameter space (p_{in}, p_{out}) of all possible straight lines passing through our data. As a pre-processing step, we applied one Lloyd iteration and then used the Delaunay-averaged gradient vectors illustrated in the right panel of figure 10. The heat map in figure 19 contains four dark squares situated along the diagonal from the upper left corner to the lower right corner. All points within those four squares define line segments

---

[21]In principle, we can define the wombling measure (7.1) in terms of the original gradient vectors $\tilde{\vec{G}}_\alpha$ or in terms of the Voronoi-averaged gradient vectors $\langle \vec{G} \rangle_i$ defined in (3.10). However, as argued in sections [4] and [5], the Delaunay-averaged vectors offer the best option for our purposes.
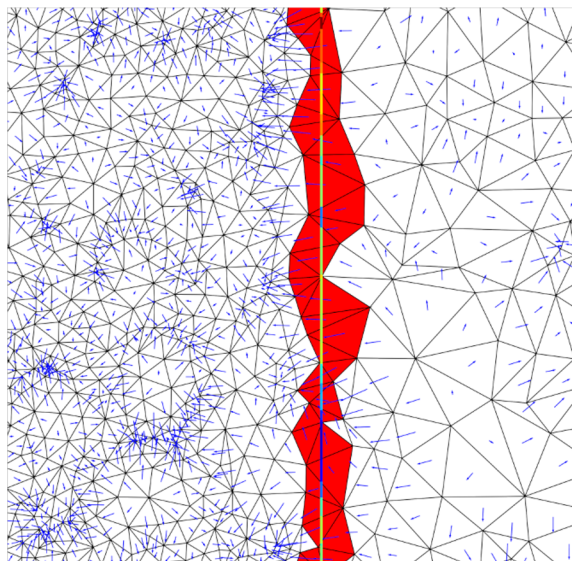
Figure 20. The line C_w with the largest wombling measure $\bar{\Gamma}[C]$ in figure 19, plotted over the Delaunay-averaged gradient vector field from the right panel of figure 10. The individual line segments $\Delta\ell_\alpha$ have been color-coded by their respective individual contributions $\langle\vec{G}\rangle_\alpha \cdot \hat{n}_C$ to the total flux. The red-shaded Delaunay cells are those which are crossed by the line C_w and are therefore included in the two summations in (7.1).

C which do not enter the field of view at all, and instead run along one of the edges of the field of view. Clearly, such line segments are irrelevant for our wombling boundary analysis, so they have been assigned $\bar{\Gamma}[C] = 0$ by default and are excluded from further consideration.

Figure 19 reveals that there is a unique line with the largest possible wombling measure — let us denote this winning line with C_w:

$$C_w \equiv \arg\max_C \left(\bar{\Gamma}[C]\right). \tag{7.2}$$

By construction, the winning line C_w is *the best* wombling boundary candidate among the set of all straight-line boundary candidates. Is it the correct wombling boundary though? According to the result from figure 19, the answer in this example is yes: C_w is found at the exact location $(0.5, 1.5)$ (or equivalently, $(2.5, 3.5)$) of the true theoretical boundary $x = 0.5$. This can be verified explicitly in figure 20, where we plot C_w overlaid on top of the Delaunay-averaged gradient vectors $\langle\vec{G}\rangle_\alpha$ from the right panel of figure 10. Figure 20 not only confirms that C_w is the correct wombling boundary, but also helps us understand why C_w was chosen by the algorithm. Note that for any given line C, the calculation of its wombling measure $\bar{\Gamma}[C]$ depends only on the Delaunay cells D_α which happen to be crossed by the line C — in figure 20 those cells are shaded in red. A careful inspection of figure 20 reveals that in the large majority of red-shaded Delaunay cells the gradient vectors are both large and (roughly) orthogonal to the line C_w, thus maximizing the average flux (7.1) through it. To better visualize this, we have color-coded the individual line segments $\Delta\ell_\alpha$ of C_w according to their individual contributions $\langle\vec{G}\rangle_\alpha \cdot \hat{n}_C$ to the total flux, with warm (cold)
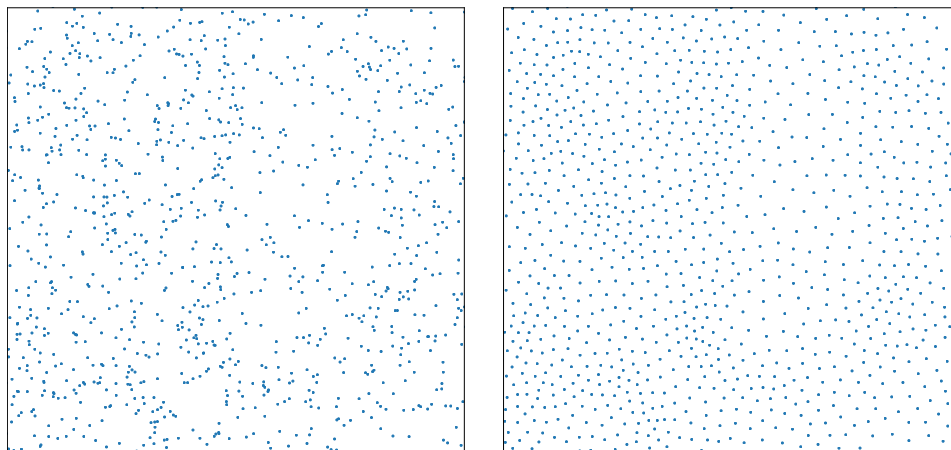
Figure 21. The point data for the straight line boundary example considered in section 7.1.2. The left panel shows the original $N = 1000$ points distributed according to (2.5) with $\rho = 1.5$. The right panel shows the same data after applying 5 Lloyd iterations.

colors corresponding to large (small) values. We see that C_w is predominantly colored with warm colors, indicating large fluxes all the way throughout. On the other hand, it is not difficult to convince oneself that this will not be true for any other randomly chosen line C crossing the field of view — the gradient vectors may happen to be relatively large and perhaps even roughly orthogonal to it purely by chance in some restricted region, but this will not occur consistently along the full length of the line as was the case with C_w. In short, the wombling procedure applied here automatically takes into account the spatial correlations of the gradient vectors along a wombling boundary.

7.1.2   An example with $\rho = 1.5$

We are now ready to tackle a more difficult case, with a smaller signal to background ratio. In this subsection we consider $N = 1000$ data points, still distributed according to (2.5), but with the much smaller value of $\rho = 1.5$. This data is shown in the left panel of figure 21. Unlike the previous example, this time the wombling boundary is not as easy to identify visually in the data. As already discussed in ref. [16], the weaker the signal, the more Lloyd steps are needed for optimal results. Correspondingly, here we apply 5 Lloyd iterations at the preprocessing stage and obtain the data shown in the right panel of figure 21, on which the subsequent wombling analysis is done.

The end result from our wombling procedure is shown in figure 22, which is the analogue of figure 19. We notice that the typical values obtained for $\bar{\Gamma}[C]$ are now much lower than what we saw in figure 19. This is to be expected due to the smaller value of ρ — the boundary is less pronounced, and the magnitudes of the gradient vectors are generally reduced as well. Despite these difficulties, the boundary is still correctly identified — notice the two bright spots in the heatmap located near $(0.5, 1.5)$ and $(2.5, 3.5)$, which is the right answer. The corresponding winning line C_w has $\bar{\Gamma}[C_w] = 0.19$ and is plotted in figure 23, where for illustration we also show a second, rather generic, line C at $(p_{\mathrm{in}}, p_{\mathrm{out}}) = (0.7, 0.8)$ which has a more typical value of the wombling parameter, $\bar{\Gamma}[C] = 0.08$. The individual line
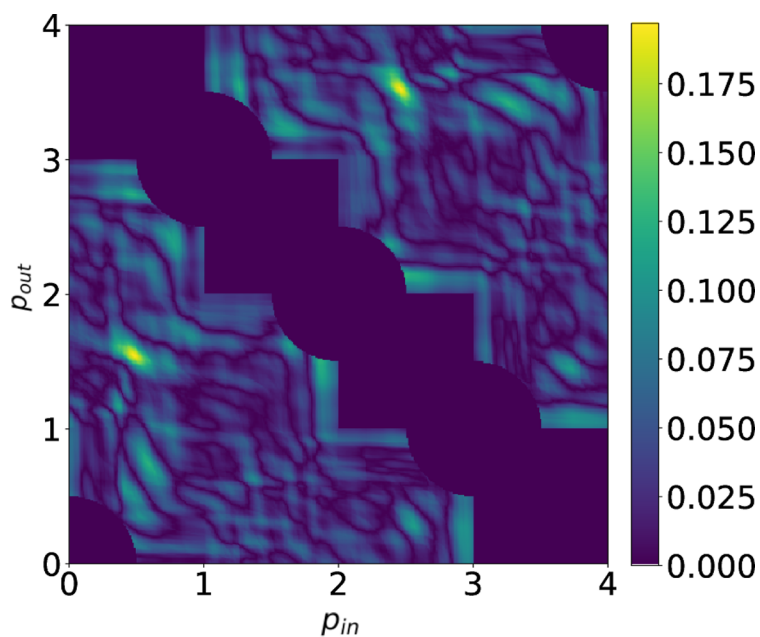
Figure 22. The same as figure 19 but for the dataset with $\rho = 1.5$ shown in the right panel of figure 21. Here we exclude from the analysis not only lines along the perimeter of the field of view but also any line of length less than 0.5; in the heatmap such lines are assigned $\bar{\Gamma}[C] = 0$ by default.
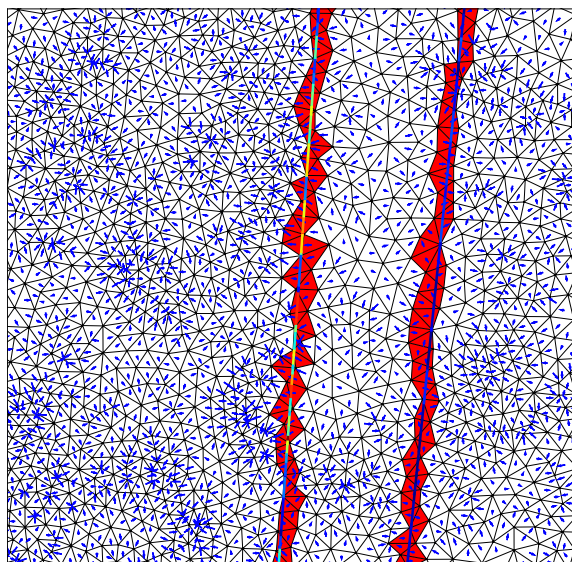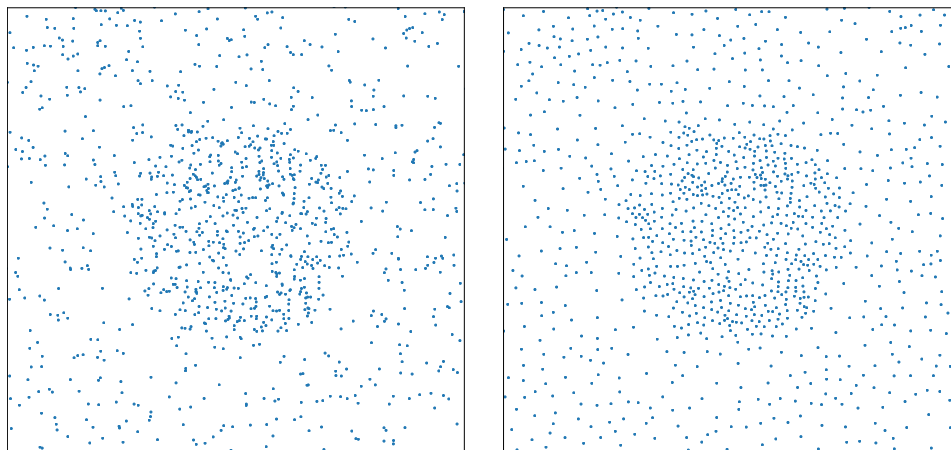

Figure 23. The same as figure 20 but for the dataset with $\rho = 1.5$ shown in the right panel in figure 21. In addition to the line C_w with the largest value of the wombling measure $\bar{\Gamma}$, for comparison we also show a generic line with a more typical value of $\bar{\Gamma}[C]$; for concreteness, we chose the line at $(p_{\mathrm{in}}, p_{\mathrm{out}}) = (0.7, 0.8)$ which happens to have $\bar{\Gamma}[C] = 0.08$.

Figure 24. The point data for the circular boundary example considered in section 7.2.1. The left panel shows the original $N = 1000$ points distributed according to (2.7) with $\rho = 5$. The right panel shows the same data after 1 Lloyd iteration.

segments of each of the two lines are color-coded similar to figure 20, i.e., proportional[22] to their individual contributions to the total flux. We see that the winner C_w is again colored with mostly warm colors, indicating large flux contributions everywhere (except for just a few spots where the gradient vectors happen to be either too small or oriented along the line), while the generic line C is colored with mostly cold colors, confirming that its wombling measure $\bar{\Gamma}$ is indeed rather small.

By comparing figures 19 and 22, one can notice that in figure 22 we have excluded from consideration not only lines which run purely along the perimeter of the field of view but also any line of length less than 0.5; this additional constraint results in the elimination of several quadrant sectors on the plot — one at the lower left corner, one at the upper right corner, and six along the diagonal running from the upper left corner to the lower right corner. Since our wombling measure is the *average* flux, short lines which literally "cut a corner" of the square field of view, can potentially pick up large gradients due to local fluctuations in the bulk, without an opportunity to cancel those fluctuations elsewhere. In other words, if we encounter two candidate lines with the same value of $\bar{\Gamma}$, and one is much longer than the other, then we would treat the longer line as the more likely wombling boundary. Thus eliminating very short lines from consideration early on would go a long way in simplifying the significance estimation procedure, see section 8 below.

7.2 A circular boundary

7.2.1 The results from a wombling analysis with straight line segments

We now proceed with the discrete version of the circular boundary example considered in section 6.2. The point dataset is shown in the left panel of figure 24 and consists of $N = 1000$ points distributed according to the probability distribution with a circular

---

[22]In order to highlight the differences between the lines segments of the two lines, we additionally apply the scaling (7.3) motivated below in section 7.2.2.
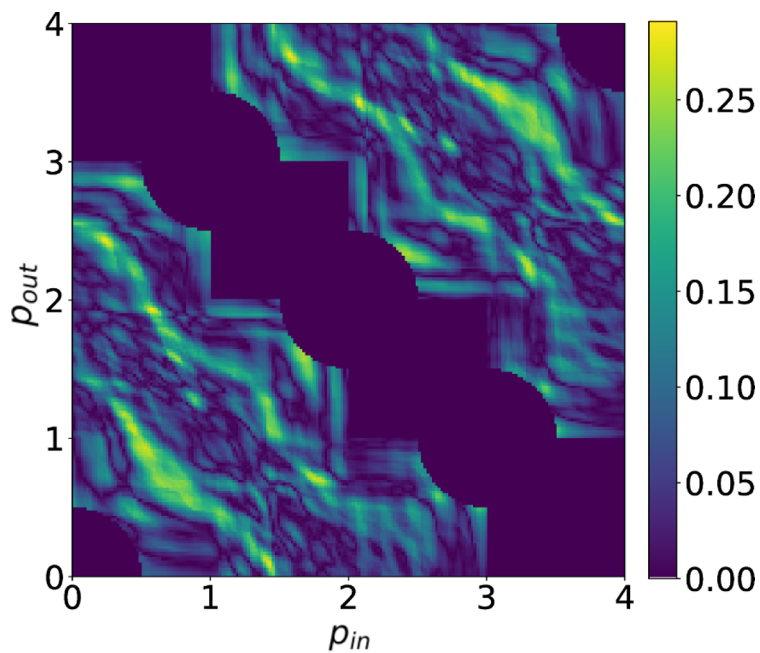
Figure 25. The same as figure 22 but for the dataset with a circular boundary shown in the right panel of figure 24.
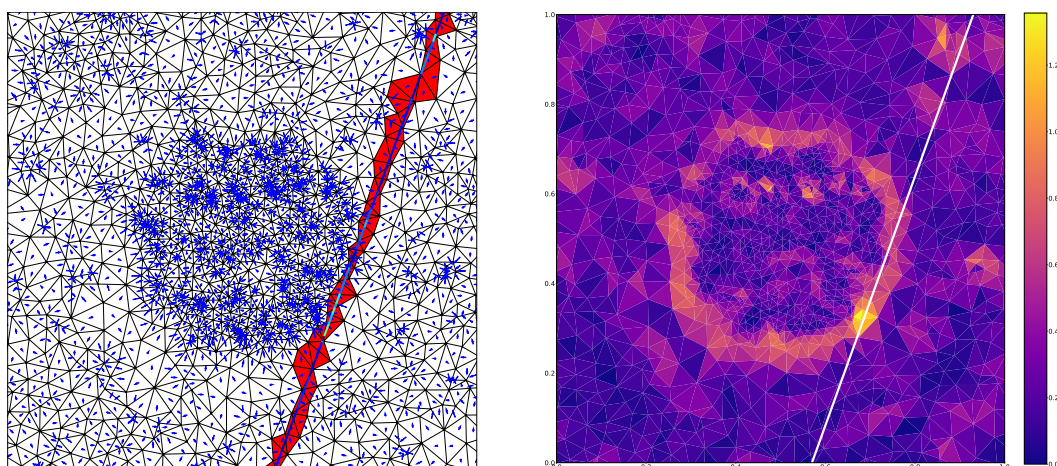

Figure 26. The line C_w with the largest wombling measure in the circular boundary example of section 7.2, overlaid on the Delaunay-averaged gradient vectors $\langle \vec{G} \rangle_\alpha$ (left panel) or on the heatmap of their magnitudes (right panel).

boundary (2.7) with $\rho = 5$. As a preprocessing step, we then apply a single Lloyd iteration, obtaining the data shown in the right panel of figure 24, on which the subsequent wombling analysis is performed.

The results from the wombling procedure are displayed in figures 25 and 26. Figure 25 is the analogue of figures 19 and 22, but for the circular boundary example considered in this subsection. Comparing to those previous figures, we notice that the largest absolute values for $\bar{\Gamma}$ obtained here are lower than those in figure 19 but higher than those in

figure 22. The former is due to the fact that we are using straight line segments to test for a curvilinear wombling boundary, and no single line segment can capture the full extent of a circular boundary, while the latter is due to the fact that the value of ρ is higher for the dataset used to produce figure 25.

Figure 25 exhibits the typical pattern observed in the continuous version of this example (the right panel of figure 18). The locations with the largest values of $|\bar{\Gamma}|$ trace out the two stripes seen in figure 18, which indicates that in the discrete version of the example it is still the line segments tangential to the circular boundary which tend to have large values of $|\bar{\Gamma}|$. The line C_w with the largest wombling measure happens to be at $(p_{\text{in}}, p_{\text{out}}) = (0.55, 1.9)$ and is plotted in figure 26, together with the set of Delaunay-averaged gradient vectors $\langle \vec{G} \rangle_\alpha$ (left panel) or the heatmap of their magnitudes (right panel). As anticipated, C_w is tangential to the circular boundary, and the largest contributions to its wombling measure indeed come from the (warm-colored) segments in the vicinity of the boundary. One should keep in mind that the particular line C_w shown in figure 26 is a very close winner among several other worthy challengers with similar values for the wombling measure — as figure 25 showed, there are several locations along the two bright stripes with similarly large values of $|\bar{\Gamma}|$.

7.2.2 Identifying the true shape of the boundary

The analysis from the previous subsection 7.2.1 demonstrated that even in the case of a curvilinear boundary, our wombling procedure produces reasonable results — it is able to identify a class of line segments, each of which already contains a portion of the true boundary. Unfortunately, none of the identified line segments is able to reproduce the full boundary all by itself. In this subsection, we shall therefore address the question of being able to globally reconstruct the wombling boundary, regardless of its shape, from the results presented so far.

In principle, there can be several different approaches to this problem.

- *Algorithmic wombling.* The standard approach is the algorithmic wombling procedure outlined in the introduction [29]. One applies a lower cut (tagging) on the magnitudes of the locally estimated gradient vectors and then suitably connects them (agglomeration). Although this approach has been subject to criticism [40], its modern implementation can perhaps benefit from some of the improvements which we have introduced here, in particular gradient rescaling (section 3.3.1), Voronoi relaxation (section 3.3.2), local averaging of gradient vectors (section 3.3.3), utilizing improved tagging variables which account for the spatial correlations among neighboring gradient vectors (section 4), etc..

- *Use the correct ansatz for the shape of the boundary.* At the cost of giving up model-independence, one could focus on a particular theory model, derive the parametric form of the expected boundary shape, and then use that parametrization to test for the presence of such boundaries in the data. This approach was proved successful in specific event topologies motivated by supersymmetry [22, 23, 56], but relies on the experimenter being able to make the correct theory model assumption.

- *Construct the envelope of the line segments with the largest wombling measures.* As shown by the results in figures 18 and 25, when probing a curvilinear boundary with straight line segments, we obtain a whole family of wombling boundary candidates which can be tagged with their relatively large values of $\bar{\Gamma}[C]$. We saw that each of these candidates is tangential to the true boundary, therefore, the task of constructing the true boundary reduces to the task of finding a planar curve which is tangential to each of the tagged straight line candidate segments at some point. The answer to this problem is precisely the envelope curve [68].

- *Identify and agglomerate "the best" line segments $\Delta\ell_\alpha$.* A specific realization of an approximate piece-wise reconstruction of the envelope curve mentioned above is offered by the following procedure. Note that our analysis not only tags straight lines C with large values of the wombling measure, but it also identifies which individual portions $\Delta\ell_\alpha$ of those lines are most likely to be tangential to a true boundary, as indicated by the rainbow-color coding of the line C_w in figures 20, 23 and 26. This suggests that instead of working with the full line C tagged by the algorithm, we can instead focus our attention on the individual elements $\Delta\ell_\alpha$ from it with the largest contributions to the average flux through C. The straightforward application of this method, however, will reintroduce sensitivity to local gradient fluctuations. In order to avoid this, we propose to rank the individual elements $\Delta\ell_\alpha$ not by their average flux $\langle\vec{G}\rangle_\alpha \cdot \hat{n}_C$, as was done in figures 20, 23 and 26, but by the *rescaled* average flux

$$\left(\langle\vec{G}\rangle_\alpha \cdot \hat{n}_C\right) \times \left|\bar{\Gamma}[C]\right|^\gamma, \tag{7.3}$$

where the power γ is a suitably chosen positive parameter, which interpolates smoothly between "complete locality" ($\gamma = 0$) and "complete globality" ($\gamma \to \infty$). The scaling by a power of $\bar{\Gamma}[C]$ in eq. (7.3) ensures that a given element $\Delta\ell_\alpha$ is judged not only by the local flux going through it, but also by its association with a suitable wombling boundary candidate line C. By increasing the power γ, we can suppress the effects of local statistical fluctuations, and in the limit of $\gamma \to \infty$ we eventually recover our previous results where one would select *all* the segments $\Delta\ell_\alpha$ belonging to the best wombling candidate line C_w. However, using finite values of γ allows us to 1) let in "the best" individual segments $\Delta\ell_\alpha$ from other candidate lines C which are not C_w, but have comparably large values of $\bar{\Gamma}[C]$, and 2) eliminate from consideration those individual segments $\Delta\ell_\alpha$ from the winning line C_w whose local flux values are too low — presumably such elements belong to C_w only because we have used the wrong ansatz for the shape of the boundary. This approach strikes the right balance between the "locality" of the flux through an individual element $\Delta\ell_\alpha$ and the "globality" (in the method of calculation) of the wombling measure $\bar{\Gamma}[C]$. The procedure is illustrated in figure 27, where we show the individual line segments $\Delta\ell_\alpha$ whose rescaled flux values (7.3) are in the top 1 percentile, after scanning the $(p_{\text{in}}, p_{\text{out}})$ parameter space on an 80×80 grid. We see that, despite using the wrong ansatz (straight lines), the circular boundary is reconstructed quite well, with only a few stragglers showing up in the bulk.
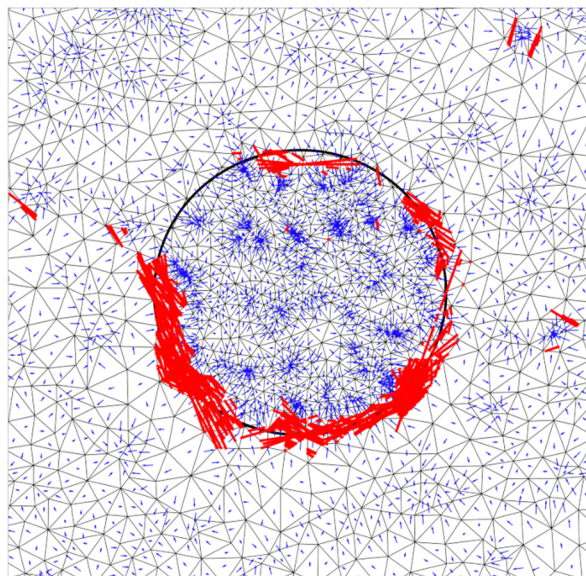
Figure 27. The individual segment selection procedure described in section 7.2.2. After scanning the $(p_{\text{in}}, p_{\text{out}})$ parameter space on an 80×80 grid and computing the wombling measures $\bar{\Gamma}[C]$, the average flux for each individual line segment $\Delta\ell_\alpha$ has been rescaled according to (7.3) with $\gamma = 4$ and the plot shows (in red) the segments with rescaled flux values in the top 1 percentile.

8 Significance estimation

In the previous sections we discussed different techniques for identifying wombling boundaries in point datasets. In applications to collider event data in high-energy physics, the presence of a wombling boundary could be indicative of new physics, if its location is in a region of phase space which is unremarkable from the point of view of the SM background. However, before claiming a discovery, one must be confident that such wombling boundaries cannot be accidentally generated by SM data alone. For this purpose, it is necessary to supplement any proposed wombling technique with a corresponding prescription for assessing the significance of any reconstructed wombling boundary. Since previous work [16, 22, 23] did not address this issue, we shall now do so using a frequentist approach.

The end result of the wombling method described in section 7 was the selection of the best possible wombling line candidate C_w, together with its corresponding wombling measure

$$\bar{\Gamma}_w \equiv \bar{\Gamma}[C_w]. \tag{8.1}$$

In order to use $|\bar{\Gamma}_w|$ as our test statistic, we need to know its distribution under the pure-background hypothesis. For this purpose, we generate a number of pseudo-experiments where the point data is generated from the pure-background distribution (2.2), and for each pseudo-experiment, we repeat the wombling analysis from the previous section. A typical result from one such pseudo-experiment is shown in figure 28, where, in analogy to figures 19, 22 and 25, we show a heat map of $|\bar{\Gamma}[C]|$ in the $(p_{\text{in}}, p_{\text{out}})$ parameter space. We see that, as expected, in the absence of a real signal the typical values for the wombling measure are relatively low almost everywhere in the $(p_{\text{in}}, p_{\text{out}})$ parameter space, except
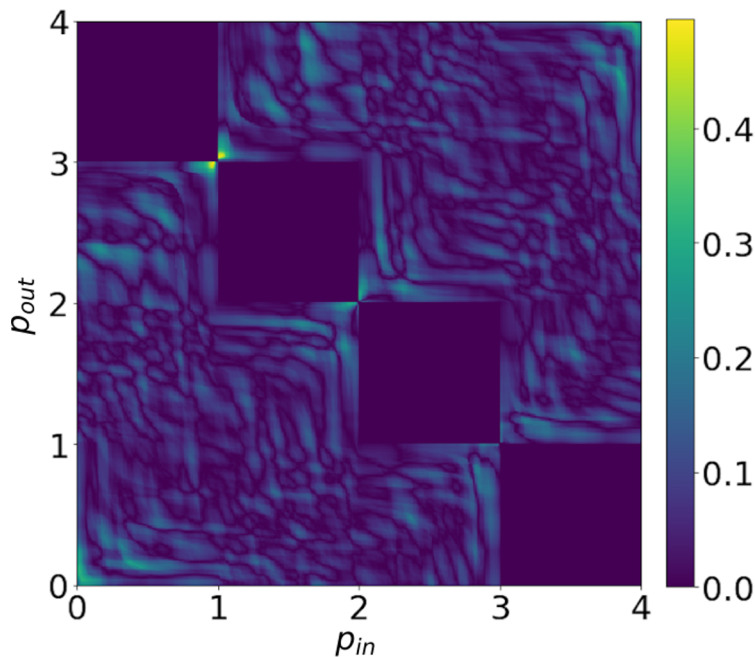
Figure 28. A heat map of wombling measures $|\bar{\Gamma}[C]|$ analogous to figures 19, 22 and 25, but for a typical background pseudo-experiment where the point data is generated from the pure-background distribution (2.2) and then a single Lloyd iteration is applied.

at one very special location near $(1,3)$. It is not difficult to realize that this location corresponds to very short candidate lines C which "cut" the lower right corner of our field of view. We already alluded to this problem at the very end of section 7.1.2, and our proposed solution was simply to exclude such very short lines from consideration.[23] Therefore, when we derive the $|\bar{\Gamma}_w|$ distributions below, we shall apply a minimum cut on the allowed length of any candidate line C. In order to make sure this problem does not reappear, we shall conservatively increase our previous minimum length cut from 0.5 to $\sqrt{2}/2$, which is the length of the line connecting the midpoints of any two neighboring edges of our field of view.

With those preliminaries, we are ready to compare the distributions of our test statistic $|\bar{\Gamma}_w|$ for signal and background. The blue histogram in figure 29 shows the $|\bar{\Gamma}_w|$ distribution for 100 pure-background pseudo-experiments with $N = 1000$ points each, where the data was generated from the pure-background distribution (2.2) as in the left panel of figure 1. At the preprocessing stage, we increased the number of Lloyd iterations to 10, since we shall be interested in the case of relatively weak signals (see the related discussion in section 7.1.2 and ref. [16]). Following the procedure of section 7, we then computed the wombling measure $\bar{\Gamma}[C]$ on an 80×80 grid in the $(p_{\text{in}}, p_{\text{out}})$ parameter space and the best wombling line C_w (of minimum length $\sqrt{2}/2$) was identified and its wombling measure (8.1) was entered in the histogram. The resulting distribution is relatively concentrated around

---

[23]Another possibility could be to use the unnormalized wombling measure (6.1). We shall leave this option open for a future investigation [69].
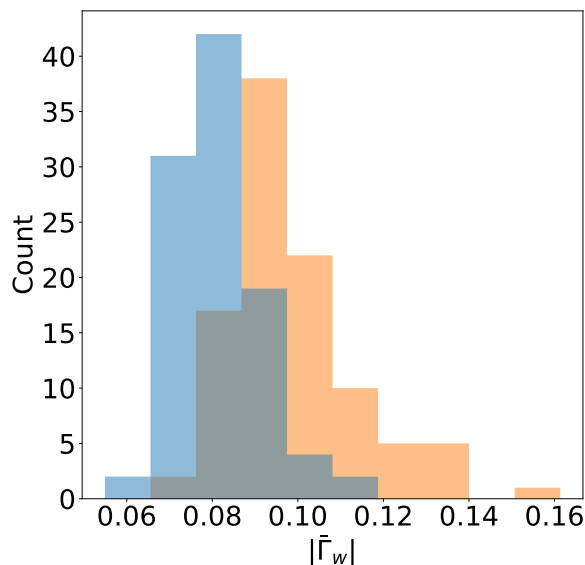
Figure 29. Distributions of $|\bar{\Gamma}_w|$ for two populations of 100 pseudo-experiments each, and with $N = 1000$ points per pseudo-experiment. The blue histogram corresponds to data generated from the pure-background distribution (2.2) as in the left panel of figure 1. The orange histogram represents pseudo-experiments with added signal with a straight-line boundary as in the middle panel of figure 1, with the points sampled from (2.5) with $\rho = 1.5$. In either case, the wombling measure $\bar{\Gamma}[C]$ was computed throughout the $(p_{\mathrm{in}}, p_{\mathrm{out}})$ parameter space on an 80×80 grid and the best wombling line C_w of minimum length $\sqrt{2}/2$ was identified following the procedure of section 7.
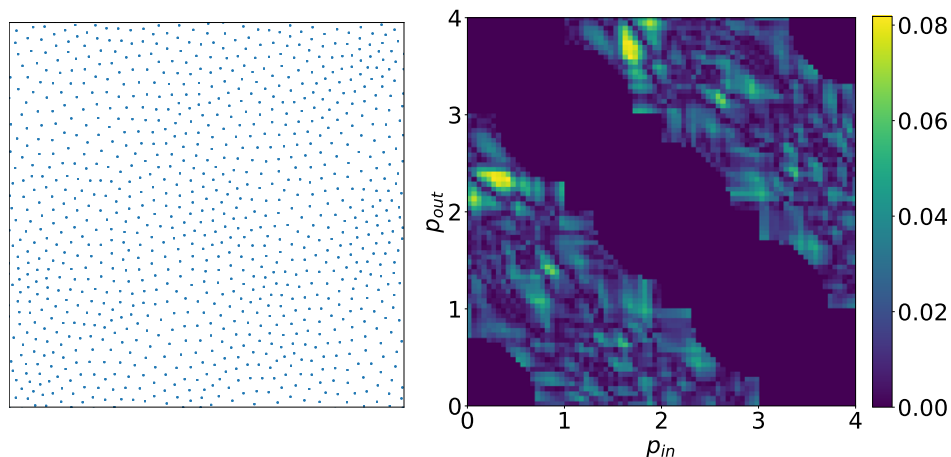
Figure 30. The point data (left panel) and the $|\bar{\Gamma}[C]|$ heatmap (right panel) for a representative pure-background pseudo-experiment entering the blue histogram in figure 29.

a mean of 0.08, and extends up to 0.11, which sets the lower limit on the target range for signal detection. For illustration, in figure 30 we show results for one typical pure-background pseudo-experiment whose value for $\bar{\Gamma}_w$ is equal to the mean of the distribution shown in figure 29.

Given the background distribution from figure 29, we can now assess what types of signals might be discoverable. Obviously, the larger the signal component, the more pro-

nounced the wombling boundary. In our conventions, the signal strength was parametrized by the ρ parameter. For example, in section 7.1.1 we saw that for $\rho = 5$ we obtained $\bar{\Gamma}_w \sim 0.7$, while the weaker signal with $\rho = 1.5$ in section 7.1.2 resulted in only $\bar{\Gamma}_w \sim 0.19$ (note that in section 7.1.2 the data was preprocessed with only 5 Lloyd steps; adding 5 additional steps as was done in figure 29 would further reduce the value of $\bar{\Gamma}_w$ slightly). Given the pure-background distribution in figure 29, it is clear that in both of those examples the observed effect could not have been attributed to a background fluctuation and would represent a discovery. At the same time, a careful inspection of the left panel in figure 21 shows that the $\rho = 1.5$ example of section 7.1.2 was rather "lucky" due to fortuitous fluctuations in the data near the theoretical boundary. In order to estimate the prospects for a more typical signal scenario, we simulate 100 pseudo-experiments with $N = 1000$ data points each, generated from the distribution (2.5) with $\rho = 1.5$. The corresponding distribution of the test statistic $|\bar{\Gamma}_w|$ for those signal pseudo-experiments is shown with the orange histogram in figure 29. We see that, on average, the values of $|\bar{\Gamma}_w|$ are larger in the presence of a signal — the mean of the orange histogram is shifted to 0.10. Comparing the tails of the two distributions, we find that in 40% of the cases, the signal is discoverable at 2 sigma and in 16% of the cases it is discoverable at 3 sigma. These prospects can probably be further improved by optimizing the different aspects of our wombling algorithm, but such an optimization is outside the scope of this paper.

9 Conclusions and outlook

In this paper we reviewed and refined the existing procedures for identifying wombling boundaries in point datasets. Our interest in this topic stems from the fact that high energy physics collider data can be viewed as point data in the relevant phase space of the final state signature. For better visual illustration, we considered point data examples in two dimensions, but our technique can be readily generalized to higher dimensional data. We proposed several modifications to the standard algorithm which lead to improved detection efficiency and significance:

- We advocated the use of the Delaunay triangulation of the data instead of (or perhaps in addition to) the Voronoi tessellation of the data. We argued that the Delaunay tessellation is the natural framework for computing the local gradient vectors which is the first and most important step of any wombling algorithm.

- We considered three different techniques for reducing the effect of statistical fluctuations:

  1. Rescaling the local gradient vectors as in (3.8), see section 3.3.1.
  2. Applying Voronoi relaxation of the data via several Lloyd iterations as a preprocessing step, see section 3.3.2.
  3. Local averaging of the gradient vectors, which can be performed either at the level of a Voronoi cell (3.9) or at the level of a Delaunay triangle (3.10), see section 3.3.3.

- We studied new tagging variables (dot products of neighboring averaged gradient vectors) for selecting elements of the tessellation marking the location of a wombling boundary. In section 4 we showed that the new variables have improved selection efficiency, since they take into account the spatial correlations among neighboring gradient vectors along the wombling boundary. In section 5 we pointed out that the new variables additionally can be used to naturally connect the tagged elements into continuous boundaries.

- In sections 6 and 7 we explored the idea of refs. [40, 63] to rank wombling boundary candidates C by a global wombling measure, e.g., $\Gamma[C]$ from (6.1) or $\bar{\Gamma}[C]$ from (6.2). On the basis of several toy examples we showed that this approach is successful in identifying the correct boundary, and with a slight modification (7.3) can be used even when the shape of the boundary is different from the assumed ansatz, see section 7.2.2.

- In section 8 we showed how one can estimate the statistical significance of any detected wombling boundary using a frequentist approach.

The present study complements and further expands the work of refs. [16, 22, 23, 25, 56] in an interesting direction which, while popular in other fields of science, is still rather new to the field of high energy physics. We believe that our investigations here are only scratching the surface of what could be a very promising research thrust. In particular, the approach of treating high energy collider data as point data and studying its geometric properties is complementary to the existing binning techniques and in the long run could prove to be more suitable to the application of modern machine learning techniques [70, 71].

Acknowledgments

A Studies of non-uniform background distributions

The illustrative examples in the main body of the paper so far have used data generated from a uniform background distribution (2.2). In most applications outside collider physics this is a valid assumption, since the density in the bulk is at most a slowly varying function. However, in particle physics one often has to face backgrounds which are steeply falling functions parametrized by power laws or exponents. For completeness, in this appendix we shall relax the uniformity assumption about the background and shall consider two other typical situations, namely, when the background distribution is given by a (linear) power law (section A.1) or an exponential (section A.2).
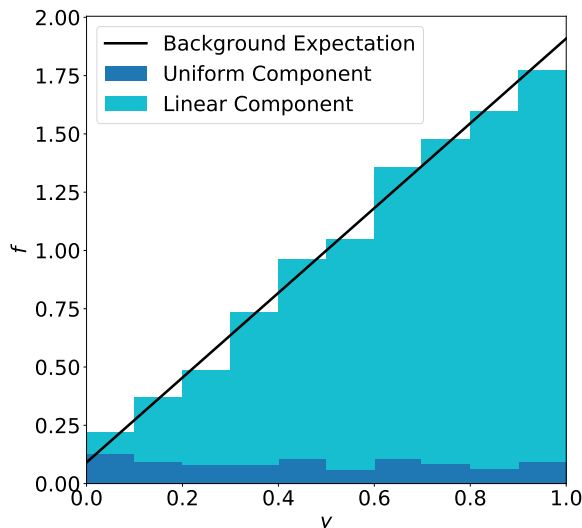
Figure 31. The background distribution (A.1) for the example with linearly increasing background considered in section A.1, for $N_U/N_L = 0.1$.

A.1 Example with linearly increasing background

In this section we reconsider our circular boundary example from section 7.2, only this time we trade the uniform background distribution (2.2) for a linearly increasing function

$$f(x,y) = \frac{N_U/N_L + 2y}{N_U/N_L + 1}. \tag{A.1}$$

Due to the rotational symmetry, without any loss of generality we can take the function f to increase in the positive y direction, as shown in eq. (A.1). The black solid line in figure 31 illustrates the y-dependence of this background distribution. Note that since $f(x, y = 0) > 0$, we can think of (A.1) as being made of two components: $f = f_U + f_L$, where the uniform component

$$f_U(x,y) = \frac{1}{1 + N_L/N_U},$$

illustrated by the dark-shaded blue histogram in figure 31, contains a total of N_U events, while the linear component

$$f_L(x,y) = \frac{2y}{1 + N_U/N_L},$$

depicted by the light-shaded cyan histogram in figure 31, contains a total of N_L events. In what follows (as well as in figure 31) we fix $N_U/N_L = 0.1$.

To this background distribution (sampled with $N_U + N_L = 1500$ points) we add a uniform circular signal as before, with total number of signal points $N_S = 450$. The resulting point dataset is shown in the left panel of figure 32. In principle, we could now run our previous analysis from section 7.2 directly on this dataset, and the result is depicted in the left panel of figure 33, which is the analogue of figure 27 for this case. In producing figure 33, we applied 7 Lloyd steps and scanned on a coarser 40×40 grid, which results
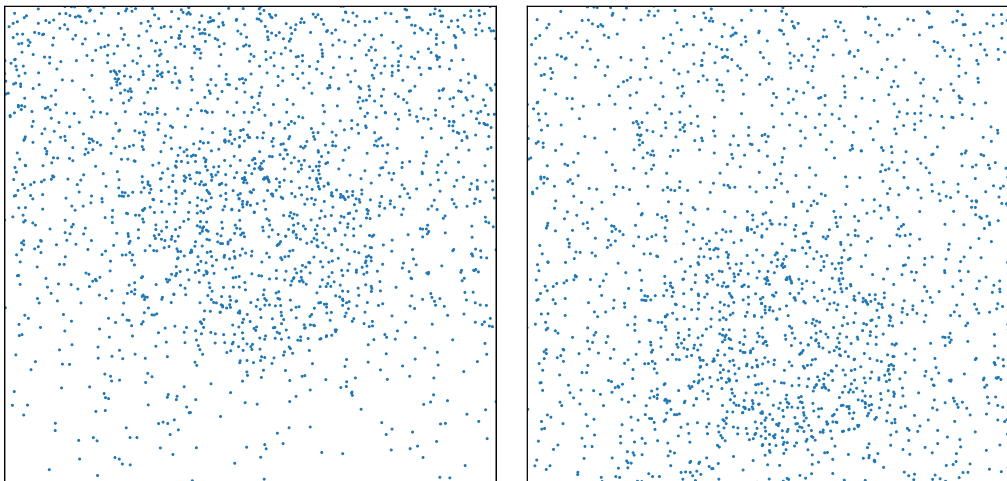
Figure 32. The point data for the example with linearly increasing background considered in section A.1 before (left panel) and after (right panel) the probability integral transform (A.2).
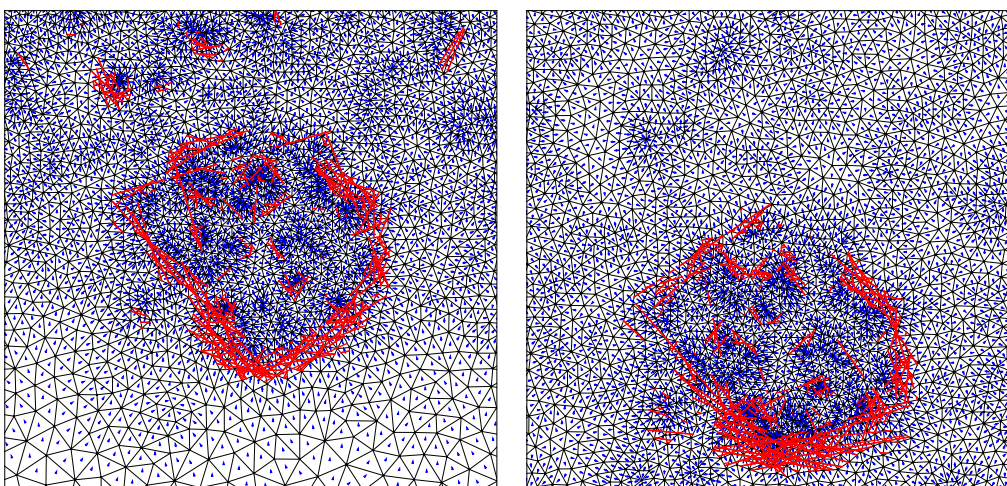

Figure 33. The same as figure 27, but for the point data in figure 32. Here we scanned on a coarser 40×40 grid after applying 7 Lloyd steps. The red lines have rescaled flux values (7.3) in the top 4 percentile (with $\gamma = 1$).

in 4 times fewer lines compared to figure 27 (which was made on an 80×80 grid). Correspondingly, to make a fair comparison with figure 27, we plotted the lines within the top 4 percentile of rescaled flux values (as opposed to the top 1 percentile shown in figure 27).

The result in the left panel of figure 33 demonstrates that our method is able to successfully identify the boundary in the presence of non-uniform background as well. However, we also notice that there are a few isolated line segments being picked up which are located in the background region near the top portion of the plot, where the background is large and the statistical fluctuations are creating relatively large local gradients. We therefore try an alternative strategy, where we first perform a change of variables designed to flatten the background distribution [72], after which we perform the wombling analysis on the resulting dataset. In cases where the background distribution is known analytically, e.g.,

as in eq. (A.1), the required transformation is simply the probability integral transform, which in our case reads

$$y \to \frac{(y+b)^2 - b^2}{2b+1},$$ (A.2)

where

$$b = \frac{N_U}{2N_L}.$$

Note that the transformation (A.2) preserves the location of the points at $y = 0$ and $y = 1$, i.e., it is a map of $[0,1] \to [0,1]$. After rescaling the y values of our original dataset, we obtain the data shown in the right panel of figure 32 — note how the density within the background region appears much more uniform, since by construction it is sampled from a uniform distribution. At the same time, the signal region has shifted down, but its boundary is still clearly defined. Now, performing our wombling analysis on the dataset in the right panel of figure 32, we obtain the result in the right panel of figure 33. We see that the boundary is again identified, this time somewhat more cleanly, since there are fewer spurious line segments in the bulk of the background region. In summary, the two panels in figure 33 show that in the case of known non-uniform backgrounds, both approaches are viable, and the choice of which one to use can be left to the individual user.

A.2 Example with exponentially increasing background

In this section we shall consider another example with non-uniform background, namely a background distribution given by

$$f(x,y) = \frac{3\,e^{3x}}{e^3 - 1}.$$ (A.3)

This function is illustrated in figure 34 with the black solid line, together with a histogram of the simulated background data points. For comparison, we also show the linearly increasing background distribution from the previous section (the black dashed line).

We shall now repeat our wombling analysis exercise, but with an added twist, to make it more challenging. First, we shall keep the same number of background points (1500) but we shall weaken our signal down to 300 points only. Second, we shall *not* assume that we know the exact analytical form (A.3) of the background distribution, so that we cannot do the background flattening exactly. Instead, we shall attempt to flatten the background using "the wrong" linear relation (A.1), since it resembles the actual (exponential) background — compare the solid and dashed lines in figure 34.

The left panel in figure 35 shows the original point dataset, while the right panel in figure 35 shows the same data after rescaling the y values with the "wrong" transformation (A.2) corresponding to a linearly increasing background. After performing our wombling analysis on these two datasets, we obtain the results shown in the respective panels of figure 36. We see that even in this more challenging exercise, the boundary is still being identified properly in the right panel, where we have applied an approximate flattening transformation. In the left panel, on the other hand, the method is still doing its job — it is finding the regions with largest gradients, which in this case are in the background region, due to the exponential behavior.
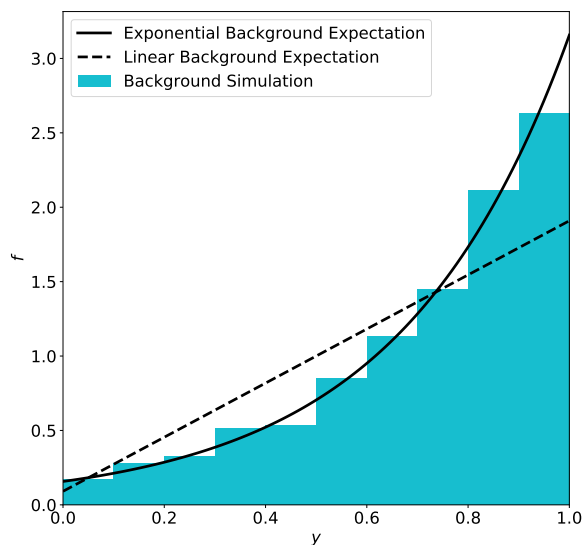
Figure 34. The unit-normalized background distribution ([A.3](#)) (solid line) and the corresponding simulated data (histogram) used for the example considered in section [A.2](#). For comparison, the dashed line shows the linear background distribution used for the example in section [A.1](#).
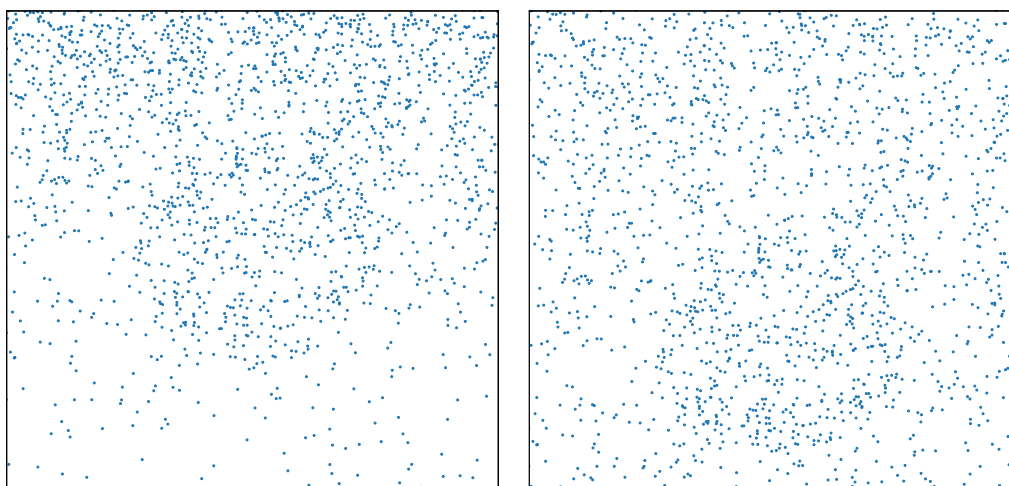

Figure 35. The same as figure [32](#), but for the example with exponentially increasing background considered in section [A.2](#). Here we use a weaker signal, with $N_S = 300$ points, while keeping the same number of background points (1500). Note that the y values in the right panel have been rescaled with the "wrong" transformation ([A.2](#)) corresponding to a linearly increasing background.
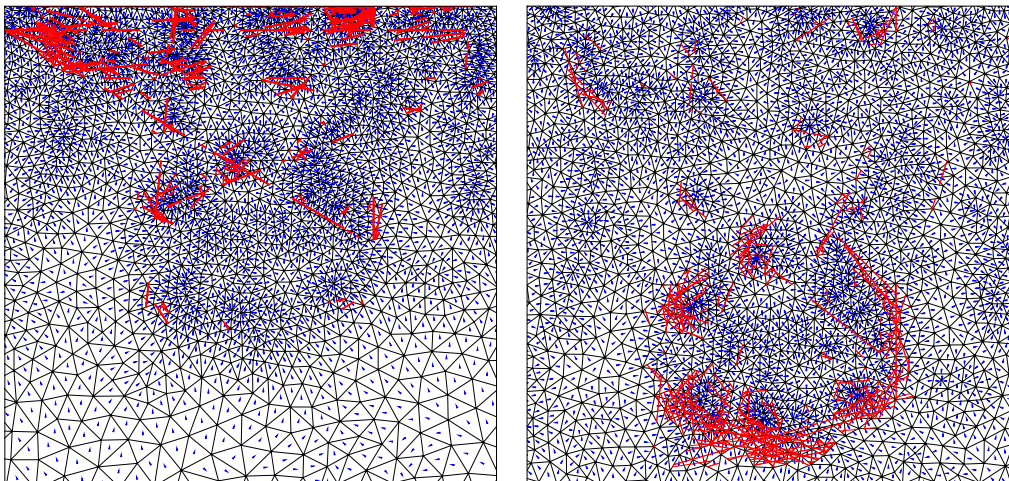
Figure 36. The same as figure 33, but for the example with exponentially increasing background considered in section A.2.

Open Access. This article is distributed under the terms of the Creative Commons Attribution License ([CC-BY 4.0](#)), which permits any use, distribution and reproduction in any medium, provided the original author(s) and source are credited.

References

[1] E.M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, *JHEP* **10** (2017) 174 [`arXiv:1708.02949`] [ɪɴSPIRE].

[2] J.A. Aguilar-Saavedra, J.H. Collins and R.K. Mishra, *A generic anti-QCD jet tagger*, *JHEP* **11** (2017) 163 [`arXiv:1709.01087`] [ɪɴSPIRE].

[3] J.H. Collins, K. Howe and B. Nachman, *Anomaly Detection for Resonant New Physics with Machine Learning*, *Phys. Rev. Lett.* **121** (2018) 241803 [`arXiv:1805.02664`] [ɪɴSPIRE].

[4] A. De Simone and T. Jacques, *Guiding New Physics Searches with Unsupervised Learning*, *Eur. Phys. J. C* **79** (2019) 289 [`arXiv:1807.06038`] [ɪɴSPIRE].

[5] J. Hajer, Y.-Y. Li, T. Liu and H. Wang, *Novelty Detection Meets Collider Physics*, *Phys. Rev. D* **101** (2020) 076015 [`arXiv:1807.10261`] [ɪɴSPIRE].

[6] T. Heimel, G. Kasieczka, T. Plehn and J.M. Thompson, *QCD or What?*, *SciPost Phys.* **6** (2019) 030 [`arXiv:1808.08979`] [ɪɴSPIRE].

[7] M. Farina, Y. Nakai and D. Shih, *Searching for New Physics with Deep Autoencoders*, *Phys. Rev. D* **101** (2020) 075021 [`arXiv:1808.08992`] [ɪɴSPIRE].

[8] A. Casa and G. Menardi, *Nonparametric semisupervised classification for signal detection in high energy physics*, `arXiv:1809.02977` [ɪɴSPIRE].

[9] O. Cerri, T.Q. Nguyen, M. Pierini, M. Spiropulu and J.-R. Vlimant, *Variational Autoencoders for New Physics Mining at the Large Hadron Collider*, *JHEP* **05** (2019) 036 [`arXiv:1811.10276`] [ɪɴSPIRE].

[10] J.H. Collins, K. Howe and B. Nachman, *Extending the search for new resonances with machine learning*, *Phys. Rev. D* **99** (2019) 014038 [`arXiv:1902.02634`] [ɪɴSPIRE].

[11] T.S. Roy and A.H. Vijay, *A robust anomaly finder based on autoencoders*, `arXiv:1903.02032` [ɪɴSPIRE].

[12] B.M. Dillon, D.A. Faroughy and J.F. Kamenik, *Uncovering latent jet substructure*, *Phys. Rev. D* **100** (2019) 056002 [`arXiv:1904.04200`] [ɪɴSPIRE].

[13] A. Blance, M. Spannowsky and P. Waite, *Adversarially-trained autoencoders for robust unsupervised new physics searches*, *JHEP* **10** (2019) 047 [`arXiv:1905.10384`] [ɪɴSPIRE].

[14] A. Mullin, H. Pacey, M. Parker, M. White and S. Williams, *Does SUSY have friends? A new approach for LHC event analysis*, `arXiv:1912.10625` [ɪɴSPIRE].

[15] B. Nachman and D. Shih, *Anomaly Detection with Density Estimation*, *Phys. Rev. D* **101** (2020) 075042 [`arXiv:2001.04990`] [ɪɴSPIRE].

[16] D. Debnath, J.S. Gainer, D. Kim and K.T. Matchev, *Edge Detecting New Physics the Voronoi Way*, *EPL* **114** (2016) 41001 [`arXiv:1506.04141`] [ɪɴSPIRE].

[17] D. Debnath, J.S. Gainer, D. Kim and K.T. Matchev, *Discovering New Physics with Voronoi Tessellations*, in *3rd Large Hadron Collider Physics Conference*, pp. 732–737 (2015) [`arXiv:1511.02724`] [ɪɴSPIRE].

[18] I.-W. Kim, *Algebraic Singularity Method for Mass Measurement with Missing Energy*, *Phys. Rev. Lett.* **104** (2010) 081601 [`arXiv:0910.1149`] [ɪɴSPIRE].

[19] A. Rujula and A. Galindo, *Measuring the W-Boson mass at a hadron collider: a study of phase-space singularity methods*, *JHEP* **08** (2011) 023 [`arXiv:1106.0396`] [ɪɴSPIRE].

[20] A. De Rujula and A. Galindo, *Singular ways to search for the Higgs boson*, *JHEP* **06** (2012) 091 [`arXiv:1202.2552`] [ɪɴSPIRE].

[21] P. Agrawal, C. Kilic, C. White and J.-H. Yu, *Improved Mass Measurement Using the Boundary of Many-Body Phase Space*, *Phys. Rev. D* **89** (2014) 015021 [`arXiv:1308.6560`] [ɪɴSPIRE].

[22] D. Debnath, J.S. Gainer, C. Kilic, D. Kim, K.T. Matchev and Y.-P. Yang, *Identifying Phase Space Boundaries with Voronoi Tessellations*, *Eur. Phys. J. C* **76** (2016) 645 [`arXiv:1606.02721`] [ɪɴSPIRE].

[23] D. Debnath, J.S. Gainer, C. Kilic, D. Kim, K.T. Matchev and Y.-P. Yang, *Detecting kinematic boundary surfaces in phase space: particle mass measurements in SUSY-like events*, *JHEP* **06** (2017) 092 [`arXiv:1611.04487`] [ɪɴSPIRE].

[24] B. Altunkaynak, C. Kilic and M.D. Klimek, *Multidimensional phase space methods for mass measurements and decay topology determination*, *Eur. Phys. J. C* **77** (2017) 61 [`arXiv:1611.09764`] [ɪɴSPIRE].

[25] D. Debnath, J.S. Gainer, C. Kilic, D. Kim, K.T. Matchev and Y.-P. Yang, *Enhancing the discovery prospects for SUSY-like decays with a forgotten kinematic variable*, *JHEP* **05** (2019) 008 [`arXiv:1809.04517`] [ɪɴSPIRE].

[26] D. Kim, K.T. Matchev and P. Shyamsundar, *Kinematic Focus Point Method for Particle Mass Measurements in Missing Energy Events*, *JHEP* **10** (2019) 154 [`arXiv:1906.02821`] [ɪɴSPIRE].

[27] K.T. Matchev and P. Shyamsundar, *Singularity Variables for Missing Energy Event Kinematics*, *JHEP* **04** (2020) 027 [`arXiv:1911.01913`] [ɪɴSPIRE].

[28] W.H. Womble, *Differential Systematics*, *Science* **114** (1951) 315.

[29] M.R.T. Dale and M.-J. Fortin, *Spatial Analysis: A Guide for Ecologists*, Cambridge University Press (2014).

[30] G. Barbujani, N.L. Oden and R.R. Sokal, *Detecting Regions of Abrupt Change in Maps of Biological Variables*, *Syst. Biol.* **38** (1989) 376.

[31] G. Barbujani, G.M. Jacquez and L. Ligi, *Diversity of some gene frequencies in European and Asian populations. V. Steep multilocus clines*, Am. J. Hum. Genet. **47** (1990) 867.

[32] N.L. Oden R.R. Sokal, M.-J. Fortin and H. Goebl, *Categorical Wombling: Detecting Regions of Significant Change in Spatially Located Categorical Variables*, *Geogr. Anal.* **25** (1993) 315.

[33] J.P. Bocquet-Appel and J.N. Bacro, *Generalized Wombling*, *Syst. Biol.* **43** (1994) 442.

[34] M.-J. Fortin, *Edge Detection Algorithms for Two-Dimensional Ecological data*, *Ecology* **75** (1994) 956.

[35] M.-J. Fortin, *Effects of Data Types on Vegetation Boundary Delineation*, *Can. J. Forest Res.* **27** (1997) 1851.

[36] G.M. Jacquez, *The map comparison problem: tests for the overlap of geographic boundaries*, *Stat. Med.* **14** (1995) 2343.

[37] G.M. Jacquez and D.A. Greiling, *Geographic boundaries in breast, lung and colorectal cancers in relation to exposure to air toxics in Long Island, New York*, *Int. J. Health Geogr.* **2** (2003) 4.

[38] J.F. Gleyze, J.N. Bacro and D. Allard, *Detecting Regions of Abrupt Change: Wombling Procedure and Statistical Significance*, in *geoENV III — Geostatistics for Environmental Applications. Quantitative Geology and Geostatistics* P. Monestiez, D. Allard, R. Froidevaux eds., vol 11, Springer, Dordrecht (2001) [DOI].

[39] M.-J. Fortin and P. Drapeau, *Delineation of Ecological Boundaries: Comparison of Approaches and Significance Tests*, *Oikos* **72** (1995) 323.

[40] S. Banerjee and A.E. Gelfand, *Bayesian Wombling*, *J. Am. Statist. Assoc.* **101** (2006) 1487.

[41] G.M. Jacquez, S. Maruca and M.-J. Fortin, *From fields to objects: A review of geographic boundary analysis*, *J. Geogr. Syst.* **2** (2000) 221.

[42] S. Okabe, B. Boots and K. Sugihara, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, John Wiley & Sons (1992).

[43] M. Cacciari, G.P. Salam and G. Soyez, *FastJet User Manual*, *Eur. Phys. J. C* **72** (2012) 1896 [arXiv:1111.6097] [INSPIRE].

[44] B. Knuteson and B. Padley, *Statistical challenges with massive data sets in particle physics*, hep-ex/0305064 [INSPIRE].

[45] B. Knuteson, *Systematic analysis of high-energy collider data*, *Nucl. Instrum. Meth. A* **534** (2004) 7 [hep-ex/0402029] [INSPIRE].

[46] B. Knuteson, *Systematic analysis of frontier energy collider data*, in *40th Rencontres de Moriond on QCD and High Energy Hadronic Interactions*, pp. 315–318 (2005) [hep-ex/0504041] [INSPIRE].

[47] D0 collaboration, *A quasi-model-independent search for new high p_T physics at DØ*, *Phys. Rev. Lett.* **86** (2001) 3712 [hep-ex/0011071] [INSPIRE].

[48] D0 collaboration, *Search for new physics in eμX data at DØ using SLEUTH: A quasi-model-independent search strategy for new physics*, *Phys. Rev. D* **62** (2000) 092004 [hep-ex/0006011] [INSPIRE].

[49] D0 collaboration, *A Quasi model independent search for new physics at large transverse momentum*, *Phys. Rev. D* **64** (2001) 012004 [hep-ex/0011067] [INSPIRE].

[50] H1 collaboration, *A General search for new phenomena in ep scattering at HERA*, *Phys. Lett. B* **602** (2004) 14 [hep-ex/0408044] [INSPIRE].

[51] CDF collaboration, *Model-Independent and Quasi-Model-Independent Search for New Physics at CDF*, *Phys. Rev. D* **78** (2008) 012002 [arXiv:0712.1311] [INSPIRE].

[52] CDF collaboration, *Model-Independent Global Search for New High-p_T Physics at CDF*, arXiv:0712.2534 [INSPIRE].

[53] CDF collaboration, *Global Search for New Physics with $2.0\,fb^{-1}$ at CDF*, *Phys. Rev. D* **79** (2009) 011101 [arXiv:0809.3781] [INSPIRE].

[54] M. Cappellari, *Voronoi binning: Optimal adaptive tessellations of multi-dimensional data*, arXiv:0912.1303 [INSPIRE].

[55] S.E. Stead, *Estimation of gradients from scattered data*, *Rocky Mt. J. Math.* **14** (1984) 265.

[56] D. Debnath, *Generic and Sensitive Searches for New Physics*, Ph.D. Thesis, University of Florida (2018).

[57] M.R.T. Dale and M.-J. Fortin, *From Graphs to Spatial Graphs*, *Ann. Rev. Ecol. Evol. Syst.* **41** (2010) 21.

[58] LHC New Physics Working Group collaboration, *Simplified Models for LHC New Physics Searches*, *J. Phys. G* **39** (2012) 105005 [arXiv:1105.2838] [INSPIRE].

[59] S.P. Martin, *A Supersymmetry primer*, *Adv. Ser. Direct. High Energy Phys.* **21** (2010) 1 [hep-ph/9709356] [INSPIRE].

[60] D. Costanzo and D.R. Tovey, *Supersymmetric particle mass measurement with invariant mass correlations*, *JHEP* **04** (2009) 084 [arXiv:0902.2331] [INSPIRE].

[61] D. Kim, K.T. Matchev and M. Park, *Using sorted invariant mass variables to evade combinatorial ambiguities in cascade decays*, *JHEP* **02** (2016) 129 [arXiv:1512.02222] [INSPIRE].

[62] S. Banerjee, *Spatial gradients and wombling*, in *Handbook of Spatial Statistics*, P. Diggle, M. Fuentes, A.E. Gelfand and P. Guttorp eds., Taylor and Francis, Boca Raton, FL (2010).

[63] A.E. Gelfand and S. Banerjee, *Bayesian wombling: finding rapid change in spatial maps*, *WIREs Comput. Stat.* **7** (2015) 307.

[64] K. Koufos and C.P. Dettmann, *Distribution of Cell Area in Bounded Poisson Voronoi Tessellations with Application to Secure Local Connectivity*, *J. Stat. Phys.* **176** (2019) 1296 [arXiv:1612.02375].

[65] M.L.V. Pitteway, *Computer graphics research in an academic environment*, Datafair '73 (1973).

[66] D.H. McLain, *Two dimensional interpolation from random data*, *Comput. J.* **19** (1976) 178.

[67] S.P. Lloyd, *Least squares quantization in PCM*, *IEEE Trans. Inform. Theory* **28** (1982) 129.

[68] *Envelope (Mathematics)*, n.d., https://en.wikipedia.org/wiki/Envelope_(mathematics).

[69] K.T. Matchev, A. Roman and P. Shyamsundar, *Pattern Recognition with Voronoi Tessellations*, in preparation.

[70] K. Albertsson et al., *Machine Learning in High Energy Physics Community White Paper*, *J. Phys. Conf. Ser.* **1085** (2018) 022008 [arXiv:1807.02876] [INSPIRE].

[71] D. Bourilkov, *Machine and Deep Learning Applications in Particle Physics*, *Int. J. Mod. Phys. A* **34** (2020) 1930019 [arXiv:1912.08245] [INSPIRE].

[72] D. Debnath, J.S. Gainer and K.T. Matchev, *Discoveries far from the Lamppost with Matrix Elements and Ranking*, *Phys. Lett. B* **743** (2015) 1 [arXiv:1405.5879] [INSPIRE].

JHEP12(2020)137