Contour Trees of Uncertain Terrains

Pankaj K. Agarwal
Duke University

Sayan Mukherjee
Duke University

Wuzhou Zhang
Duke University

ABSTRACT

We study contour trees of terrains, which encode the topological changes of the level set of the height value ℓ as we raise ℓ from $-\infty$ to $+\infty$ on the terrains, in the presence of uncertainty in data. We assume that the terrain is represented by a piecewise-linear height function over a planar triangulation $\mathbb{M}$, by specifying the height of each vertex. We study the case when $\mathbb{M}$ is fixed and the uncertainty lies in the height of each vertex in the triangulation, which is described by a probability distribution. We present efficient sampling-based Monte Carlo methods for estimating, with high probability, (i) the probability that two points lie on the same edge of the contour tree, within additive error; (ii) the expected distance of two points p, q and the probability that the distance of p, q is at least ℓ on the contour tree, within additive error, where the distance of p, q on a contour tree is defined to be the difference between the maximum height and the minimum height on the unique path from p to q on the contour tree. The main technical contribution of the paper is to prove that a small number of samples are sufficient to estimate these quantities. We present two applications of these algorithms, and also some experimental results to demonstrate the effectiveness of our approach.

Categories and Subject Descriptors

F.2 [**Analysis of algorithms and problem complexity**]: Nonnumerical algorithms and problems; H.3.1 [**Information storage and retrieval**]: Content analysis and indexing—*indexing methods*

Keywords

Contour trees, stochastic process, data uncertainty, Monte Carlo method

1. INTRODUCTION

Let $\mathbb{M}$ be a triangulation of $\mathbb{R}^2$, and let $h : \mathbb{M} \to \mathbb{R}$ be a continuous function, called a *height function*, that is linear within each triangle of $\mathbb{M}$. The graph of h, called a *terrain* and denoted by Σ, is a triangulated xy-monotone surface in $\mathbb{R}^3$. There is extensive work on modeling and analyzing terrain data in several disciplines including GIS, spatial data bases, computational geometry, and environmental sciences.

Given a height value ℓ, the level set of h is the set of all points in $\mathbb{M}$ whose height values are ℓ. As ℓ varies, the level set continuously deforms and its topology changes at certain heights. Level sets and their topology are often used for the analysis and visualization of terrains. The contour tree of a height function encodes the evolution of the level sets and succinctly represents the topology of all level sets, and it has found applications in a wide array of data analysis and visualization problems [4, 7, 10, 12, 20].

Most of the work in terrain analysis in GIS and computational geometry assumes that the height function is given exactly and there is no uncertainty in data. But there are several sources of uncertainty in modern terrain data, such as LiDAR data, with measurement errors and missing data being the most obvious ones. There are also other sources of uncertainty. For example, the raw LiDAR data contains points returned from various anthropogenic objects such as buildings, bridges, vehicles, and trees. Therefore classification algorithms are used to filter and classify such points before constructing a digital elevation model (DEM) for the terrain. These algorithms are not completely accurate and often attach a probability of a point being a ground point or return multiple possible heights of a point each attached with a confidence level. See for example [11] and the references therein. Another source of uncertainty in terrain data is anonymization of data because of privacy concerns. Motivated by these applications, we study contour trees for terrains under data uncertainty. We focus on the case when $\mathbb{M}$ is fixed but the height function is described as a probability distribution H. That is, instead of a single surface we have a spatial process. We refer to $(\mathbb{M}, \mathsf{H})$ as an *uncertain terrain*.

A natural question to ask in the context of uncertain terrains is: what is a contour tree of an uncertain terrain? There can be exponential number of contour tree instances, one for each instance of the height function in H, each of which appears with a low probability, so computing the most-likely contour tree will be neither useful nor efficient. We therefore focus on computing certain basic statistics on contour trees, which are motivated by applications of contour trees. In particular, we study the following two questions:

(Q1) Given two points $p, q \in \mathbb{R}^2$, what is the probability of

SIGSPATIAL'15 November 03-06, 2015, Bellevue, WA, USA
2015 ACM 978-1-4503-3967-4/15/11 ...$15.00.
http://dx.doi.org/10.1145/2820783.2820823.

p and q lying on the same edge of the contour tree?

(Q2) What is the expected distance of two points $p, q \in \mathbb{R}^2$ on the contour tree, where the distance of p, q on a contour tree is defined to be the difference between the maximum height and the minimum height on the unique path from p to q on the contour tree, as defined in [6]? Alternatively, given a value ℓ, what is the probability of the distance between p and q exceeding ℓ? This definition of distance has been used for navigation on terrains as well as for measuring similarity between two terrains [6, 8].

Our contributions. We present near-linear size data structures for answering (Q1) and (Q2) queries efficiently. These data structures follow a simple Monte-Carlo approach: fix a parameter s, choose s random height functions from H, and preprocess each of them to answer deterministic counterparts of queries (Q1) and (Q2) (determine whether p and q lie on the same edge of the contour tree, or compute the distance between p and q). Given a pair of query points $p, q \in \mathbb{R}^2$, we query all s data structures with p, q and estimate for (Q1) or (Q2) based on these results. The total size, query time, and preprocessing time of these data structures are $O(sn)$, $O(s \log n)$, and $O(sn \log n + st(n))$, respectively, where $t(n)$ is the time needed to draw a random height function from H.

The main technical challenge is to bound the value of s to ensure a desired level of accuracy. Since the number of different contour trees can be exponential in n, a standard analysis based on chernoff bounds suggests that one has to set $s = \Omega(n)$ [22]. Using sophisticated techniques from combinatorial geometry and probabilistic methods, we show that only $O(\frac{1}{\varepsilon^2} \log \frac{n}{\delta})$ samples are needed to ensure that the answer is correct within error ε for all queries with probability at least $1 - \delta$. We first prove this bound in Section 3 for the (Q1) query, and then prove in Section 4 for the (Q2) query.

Next, we show (in Section 5) that answering the above queries can be used for computing topological persistence and hydrology analysis in the presence of uncertainty. Finally, we present experimental results (in Section 6) to demonstrate the efficacy of our approach.

Related work. Efficient algorithms have been devised for computing contour trees of terrains both in the RAM model [10, 23] and in the I/O model [2], and for maintaining contour trees of dynamic terrains [1], where terrains are represented as triangulated irregular networks (TINs). The contour-tree problem under uncertainty has received some attention recently, see e.g. [17, 19, 21]. Kraus [19] studied the visualization of uncertain contour trees, where he showed how to determine (by using grayscale morphology) and visually convey the uncertainty of the elements of a contour tree, and how to combine multiple contour trees of different versions of a data set in one visualization. Mihai and Westermann [21] studied the visualization of the stability of critical points in uncertain scalar fields, where they derived measures for the likelihood of a critical point occurring around a given location. Günther et al. [17] studied *mandatory critical points* of 2D uncertain scalar fields, where a mandatory critical point is represented by a *critical component* as well a *critical interval* such that any realization of the field has at least one critical point of a given type present in the critical component and taking a value in the critical interval. The mandatory critical points can be interpreted as the *common topological denominator* of all the realizations of the uncertain data.

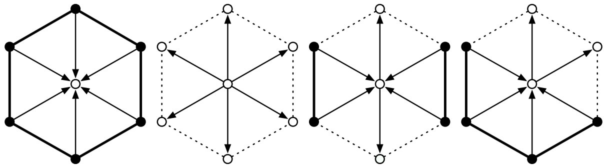


Figure 2. Maximum, minimum, saddle, and regular vertices; hollow (resp. filled) vertices in the link are lower (resp. higher).

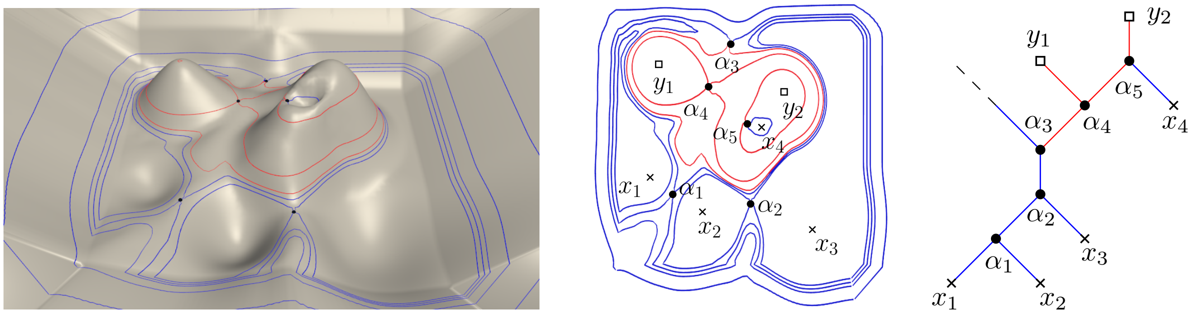
There is extensive work in spatial statistics on estimating on inferring various statistical properties of a terrain under uncertainty; see the book [5] for details. However this line of work does not focus on developing efficient algorithms. In the context of algorithms for uncertain terrains, there is some recent work on path planning [15, 16]. Gray and Evans [16] showed that finding the *optimistic* shortest path on uncertain terrains is NP-hard, where the height of each vertex is uniformly distributed in an interval and the optimistic path is the shortest path among all realizations of the terrain. Later, Gary [15] extended the hardness result to the *pessimistic*-shortest-path problem, i.e., computing the longest optimal path among all terrain instances of the uncertain terrain.

2. PRELIMINARIES

Terrains. Let $\mathbb{M} = (V, E, F)$ be a triangulation of $\mathbb{R}^2$, with vertex, edge, and face (triangle) sets V, E, and F, respectively, and let $n = |V|$. We assume that V contains a vertex v_∞ at infinity, and that each edge $\{u, v_\infty\}$ is a ray emanating from u; the triangles in $\mathbb{M}$ incident to v_∞ are unbounded. Let $h : \mathbb{M} \to \mathbb{R}$ be a *height function*. We assume that the restriction of h to each triangle of $\mathbb{M}$ is a linear map, that h approaches $-\infty$ at v_∞, and that the heights of all vertices are distinct. Given $\mathbb{M}$ and h, the graph of h, called a *terrain* and denoted by Σ_h, is an xy-monotone triangulated surface whose triangulation is induced by $\mathbb{M}$. See Fig. 1. If h is clear from the context, we denote Σ_h by Σ. The vertices, edges, and faces of Σ are in one-to-one correspondence with those of $\mathbb{M}$. With a slight abuse of terminology we refer to V, E, and F, as vertices, edges, and triangles of both Σ and $\mathbb{M}$.

Critical points. For a vertex v of $\mathbb{M}$, the *link* of v, denoted by $\mathrm{Lk}(v)$, is the cycle formed by the edges of $\mathbb{M}$ that are not incident on v but belong to the triangles incident to v. The lower (resp. upper) link of v, $\mathrm{Lk}^-(v)$ (resp. $\mathrm{Lk}^+(v)$), is the subgraph of $\mathrm{Lk}(v)$ induced by vertices u with $h(u) < h(v)$ (resp. $h(u) > h(v)$). A *minimum* (resp. *maximum*) of $\mathbb{M}$ is a vertex v for which $\mathrm{Lk}^-(v)$ (resp. $\mathrm{Lk}^+(v)$) is empty. A maximum or a minimum vertex is called an *extremal* vertex. A non-extremal vertex v is *regular* if $\mathrm{Lk}^-(v)$ (and also $\mathrm{Lk}^+(v)$) is connected, and *saddle* otherwise. A vertex that is not regular is called a *critical* vertex. See Fig. 2.

Level sets and contours. Given any value $\ell \in \mathbb{R}$, the *ℓ-level set*, the *ℓ-sublevel set*, and the *ℓ-superlevel set* of $\mathbb{M}$, denoted as $\mathbb{M}_\ell$, $\mathbb{M}_{<\ell}$, $\mathbb{M}_{>\ell}$, respectively, consist of points $x \in \mathbb{R}^2$, with $h(x) = \ell$, $h(x) < \ell$ and $h(x) > \ell$, respectively.



(a) (b) (c)

Figure 1. (a) (b) An example terrain depicted with contours through saddle vertices and showing the critical vertices of the terrain: $\alpha_1, \alpha_3, \alpha_4, \alpha_5$ are saddles. (c) The contour tree of the terrain in (a). This figure is taken from [1].

A connected component of $\mathbb{M}_\ell$ is called a *contour*. See Fig. 1. Each point $v \in \mathbb{R}^2$ is contained in exactly one contour in $\mathbb{M}_{h(v)}$, which we call *the contour of* v. The contour of a noncritical point is a simple polygonal cycle with non-empty interior. The contour of a local minimum or maximum v only consists of the single point v, and the contour of a saddle vertex v consists of two or more simple cycles with v being their only intersection point.

Contour trees. Consider raising ℓ from $-\infty$ to ∞. The contours continuously deform, but no changes happen to the topology of the level set as long as ℓ varies between two consecutive critical levels. A new contour appears as a single point at a minimum vertex, and an existing contour contracts into a single point and disappears at a maximum vertex. An existing contour splits into two new contours or two contours merge into one contour at a saddle vertex. The *contour tree* T_h of h is a tree on the critical vertices of $\mathbb{M}$ that encodes these topological changes of the level set. An edge (v, w) of T_h *represents* the contour that appears at v and disappears at w.

Formally, T_h is the quotient space in which each contour is represented by a point and connectivity is defined in terms of the quotient topology. Let $\rho : \mathbb{M} \to \mathsf{T}_h$ be the associated quotient map, which maps all points of a contour to a single point on an edge of T_h. Fix a point p in $\mathbb{M}$. If p is not a critical vertex, $\rho(p)$ lies in the relative interior of an edge in T_h; if p is an extremal vertex, $\rho(p)$ is a leaf node of T_h; and if p is a saddle vertex then $\rho(p)$ is a non-leaf node of T_h. See Fig. 1. We will use h to denote the height function on the points of T_h as well.

The preimages of points on an edge (u, v) of the contour tree is a connected planar region bounded by the contours of u and v. These regions induce a planar subdivision (also defined in [8]), which we call the *height-level map* of Σ, and denote by M_h. See Fig. 3. Note that M_h can have $\Theta(n^2)$ vertices. We write $p \sim_h q$ if $\rho(p)$ and $\rho(q)$ lie on the same edge of T_h, i.e., p, q lie in the same face of M_h. We use $\mathbf{1}(p \sim_h q)$ to denote the indicator function for $p \sim_h q$, i.e., $\mathbf{1}(p \sim_h q) = 1$ if $p \sim_h q$, and 0 otherwise.

We define the *extended height-level map* of Σ to be the planar subdivision induced by the level sets through all vertices of the triangulation, instead of the contours of critical vertices. The extended height-level map has the same worst case complexity as the height-level map.

We use Σ_h and T_h to derive a distance function in $\mathbb{R}^2$, denoted by $\mathrm{d}_h(\cdot, \cdot)$. For two points $p, q \in \mathbb{R}^2$, let $\chi(p, q)$ denote the unique path from $\rho(p)$ to $\rho(q)$ in T_h. Then

$$\mathrm{d}_h(p, q) = \max_{x \in \chi(p,q)} h(x) - \min_{x \in \chi(p,q)} h(x).$$

Intuitively, $\mathrm{d}_h(p, q)$ is the minimum height change needed to go from p to q on Σ_h. See [6, 8].

Merge trees and split trees. Analogous to the contour tree of Σ, which encodes the topological changes in Σ_ℓ as we increase ℓ from $-\infty$ to ∞, the *merge tree* (resp. *split tree*) encodes the topological changes in $\Sigma_{<\ell}$ (resp. $\Sigma_{>\ell}$). Its leaves are minima (resp. maxima) of Σ and internal nodes are (some of the) saddle vertices of Σ.

Uncertainty model. In our setup, $\mathbb{M}$ is fixed but the height function is drawn from a distribution H. We assume that the height of each vertex is drawn independently. We consider two cases. First, we assume that the height of vertex v_i, $h(v_i)$, is drawn from a discrete set $\mathsf{H}_i = \{h_i^1, \ldots, h_i^k\}$ with $\Pr[h(v_i) = h_i^j] = \gamma_i^j$, where $\gamma_i^j \in [0, 1]$ and $\sum_{j=1}^k \gamma_i^j = 1$. We say that H has *description complexity* k. For simplicity, we also assume that for any pair of distinct vertices v_ℓ, v_r, $\mathsf{H}_\ell \cap \mathsf{H}_r = \emptyset$. We also consider $h(v_i)$ being drawn from a continuous distribution defined by a probability density function (pdf) $\gamma_i : \mathbb{R} \to \mathbb{R}_{\geq 0}$; examples include the uniform and Gaussian distributions.

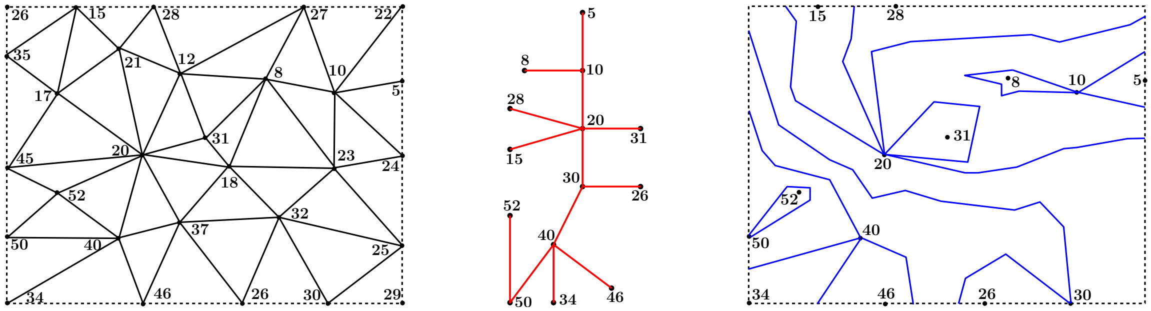
We use h to denote a random height function drawn from H. Since h is completed determined by the heights of the vertices, we will sometimes represent h as a vector $\langle h_1, \ldots, h_n \rangle$ where $h_i = h(v_i)$. We use $\gamma(h)$ to denote the probability of the outcome h, i.e.,

$$\gamma(h) = \prod_{i=1}^{n} \Pr[h(v_i) = h_i].$$

H induces distributions $\mathbf{\Sigma}_\mathsf{H}, \mathsf{T}_\mathsf{H}$ and M_H over terrains, contour trees, and heights level maps. Σ_h, T_h, and M_h are random terrain, contour tree and height-level map drawn from $\mathbf{\Sigma}_\mathsf{H}, \mathsf{T}_\mathsf{H}$ and M_H, respectively. We note that if H is a discrete distribution of description complexity k, then H has k^n different height functions and T_H and M_H can have $\Theta(k^n)$ size; see [25] for a lower bound construction.

3. PROBABILITY OF TWO POINTS LYING ON AN EDGE OF THE CONTOUR TREE

Given $\mathbb{M}$ and a distribution H on the height functions, we wish to build a data structure that can quickly compute



(a) (b) (c)

Figure 3. (a) A terrain Σ; (b) its contour tree T_h; (c) its height-level map M_h. The figure is taken from [8].

$\pi(p,q)$, the probability of p,q lying on the same edge of the contour tree. Note that $\pi(p,q) = \sum_{h\in\mathsf{H}} \gamma(h)\cdot \mathbf{1}(p \sim_h q)$. Since $|\mathsf{H}| = \Theta(k^n)$, computing $\pi(p,q)$ exactly seems hard. We describe a simple Monte-Carlo algorithm that, given two parameters $\varepsilon, \delta \in (0,1)$, computes a value $\hat{\pi}(p,q)$ such that $|\pi(p,q) - \hat{\pi}(p,q)| \le \varepsilon$ with probability at least $1-\delta$, for any $p,q \in \mathbb{R}^2$.

A Monte-Carlo algorithm. We fix a value $s \ge 1$, to be specified later. The preprocessing algorithm works in s rounds. In the j-th round, the algorithm randomly chooses the height of each vertex v of $\mathbb{M}$, denoted by $h_j(v)$, according to the distribution H. Let $h_j : \mathbb{M} \to \mathbb{R}$ be the resulting height function, let Σ_j denote the resulting terrain, let T_j denote its contour tree, and let ρ_j denote the corresponding quotient map. For each $j \le s$, using the algorithm in [8], we construct the linear-size data structure in $O(n\log n)$ time such that given two points p and q in $\mathbb{R}^2$, one can determine in $O(\log n)$ time whether $p \sim_{h_j} q$.

Given two points $p,q \in \mathbb{R}^2$, for each $j \le s$, we query the data structure to determine whether $p \sim_{h_j} q$. If the answer is yes for c instances, we return $\hat{\pi}(p,q) = c/s$.

The total size and the query time of the data structure are $O(sn)$ and $O(s\log n)$, respectively. Next we determine the value of s so that $|\hat{\pi}(p,q) - \pi(p,q)| \le \varepsilon$ for all pairs of p and q in $\mathbb{R}^2$, with probability at least $1-\delta$.

For fixed p,q, and for $j \le s$, let $\mathsf{X}_j = \mathbf{1}(p \sim_{h_j} q)$. Note that $\mathsf{X}_j \in \{0,1\}$, $\mathsf{E}[\mathsf{X}_j] = \pi(p,q)$, and $\hat{\pi}(p,q) = \frac{1}{s}\sum_{j=1}^{s}\mathsf{X}_j$. Applying the Chernoff bound [9], we obtain

$$\Pr\left[|\hat{\pi}(p,q) - \pi(p,q)| \ge \varepsilon\right] \le 2\exp(-2\varepsilon^2 s). \qquad (1)$$

Let Θ denote a set of *representative pairs* of points such that, if the inequality $|\hat{\pi}(p,q) - \pi(p,q)| \le \varepsilon$ for all pairs $(p,q) \in \Theta$, then the inequality holds for any two points p,q in $\mathbb{R}^2$. By applying the union bound to Eq. (1), the probability that there exist two points $p,q \in \mathbb{R}^2$ and $|\hat{\pi}(p,q) - \pi(p,q)| \ge \varepsilon$ is at most $2|\Theta|\exp(-2\varepsilon^2 s)$. Hence, by setting

$$s = s(\varepsilon) := \frac{1}{2\varepsilon^2}\ln\frac{2|\Theta|}{\delta}, \qquad (2)$$

$|\hat{\pi}(p,q) - \pi(p,q)| \le \varepsilon$ for all pairs of points $p,q \in \mathbb{R}^2$ with probability at least $1-\delta$.

To complete the argument, we show below (i) how to choose a random height function from H and (ii) the existence of a representative set Θ. First we consider the case when H is a discrete distribution, and then extend the argument to continuous distributions.

Discrete case. Since each vertex v has k possible values, the above algorithm can be implemented very efficiently. Each $h_j(v)$ can be selected in $O(\log k)$ time after preprocessing each $h(v)$, in $O(k)$ time, into a balanced binary tree with total weight calculated for each subtree [22]. Thus total preprocessing takes $O(s(n(\log n + \log k)) + nk) = O(nk + sn\log(nk))$ time and $O(sn)$ space, and each query takes $O(s\log n)$ time.

We now describe how to choose the representative set Θ. Let H be a discrete distribution of description complexity k on the height function, and let M_H be the distribution on the height-level map induced by H. Note that H and M_H have exponential size.

We compute the overlay $\widehat{\mathsf{M}}$ of all height-level maps in M_H, i.e., $\widehat{\mathsf{M}}$ is a planar subdivision in which two points p and q lie on the same edge or in the same face if and only if $p \sim_h q$ for all height functions h in H. Since each map in M_H is a polygonal planar subdivision, so is $\widehat{\mathsf{M}}$. For each vertex, edge, or face ϕ of $\widehat{\mathsf{M}}$, we choose an arbitrary point ξ_ϕ, and we set

$$\Theta = \{(\xi_\phi, \xi_{\phi'}) \mid \phi, \phi' \in \widehat{\mathsf{M}}\}.$$

We remark that Θ is only needed for the analysis and not for the data structure.

Lemma 3.1. *Θ is a representative set.*

Proof. Let $p,q \in \mathbb{R}^2$ be two arbitrary points, and let ϕ_p, ϕ_q be the features (vertices, edges, or faces) of $\widehat{\mathsf{M}}$ containing p and q, respectively. Then $\pi(p, \xi_{\phi_p}) = \pi(q, \xi_{\phi_q}) = 1$. Consequently, for any $h \in \mathsf{H}$, $p \sim_h q$ if and only if $\xi_{\phi_p} \sim_h \xi_{\phi_q}$, which implies that $\pi(p,q) = \pi(\xi_{\phi_p}, \xi_{\phi_q})$. Since $(\xi_{\phi_p}, \xi_{\phi_q}) \in \Theta$, the lemma follows. □

Despite the size of M_H being exponential in n, we prove below that the number of vertices, edges and faces in $\widehat{\mathsf{M}}$ is only polynomial in n and k.

Lemma 3.2. *Given a triangulation $\mathbb{M} = (V,E,F)$ in $\mathbb{R}^2$, where $n = |V|$, and a discrete distribution of description complexity k over the height functions, $\widehat{\mathsf{M}}$ has complexity $O(n^3k^8)$.*

Proof. Note that in the exact case, each triangle is crossed at most once by the contour of a saddle vertex, i.e., each triangle contributes at most one edge to the contour of a saddle vertex in M_h. There are n possible saddle vertices, each taking k possible values, giving us nk possible saddle values. There are k^3 vertex value combinations for the three vertices of a triangle, hence there are $O(nk^4)$ segments inside each triangle. The *arrangement* of these $O(nk^4)$ segments (i.e., the planar decomposition induced by these segments) has complexity $O(n^2k^8)$. Since we have n triangles and segments are disjoint between any two triangles, the complexity of the resulting planar map, which we denote by $\widetilde{\mathsf{M}}$, is $O(n^3k^8)$. Furthermore, $\widetilde{\mathsf{M}}$ is finer than $\widehat{\mathsf{M}}$, hence $\widehat{\mathsf{M}}$ also has complexity $O(n^3k^8)$. $\square$

Corollary 3.3. $|\Theta| = O(n^6k^{16})$.

Plugging in the bound on $|\Theta|$ into our above Monte-Carlo algorithm, we obtain the following result.

Theorem 3.4. *Given a triangulation $\mathbb{M} = (V, E, F)$ in $\mathbb{R}^2$, where $n = |V|$, a discrete distribution of description complexity k over the height functions, and two parameters $\varepsilon, \delta \in (0, 1)$, a data structure of size $O((n/\varepsilon^2)\log(nk/\delta))$ can be constructed in time $O(nk + (n/\varepsilon^2)\log(nk)\log(nk/\delta))$ that for any two points $p, q \in \mathbb{R}^2$, in $O((1/\varepsilon^2)\log(nk/\delta)\log n)$ time, returns a value $\hat{\pi}(p, q)$ such that $|\hat{\pi}(p, q) - \pi(p, q)| \le \varepsilon$ with probability at least $1 - \delta$.*

Remarks. If two points in $\mathbb{R}^2$ map to the same edge of the contour tree, then they also map to the same edge of the merge or split tree. Therefore, a similar analysis can be used to prove that Theorem 3.4 also holds for merge and split trees.

Continuous case. There are two technical issues in extending this technique and analysis to continuous distributions. First, we sample the height of each vertex v from its continuous distribution $h(v)$. Herein we assume the representation of the pdf is such that this can be done in constant time for each $h(v)$.

Second, we need to bound the number of distinct queries that need to be considered to apply the union bound as we did above. Since $\pi(p, q)$ may vary continuously with the locations of p, q, unlike the discrete case, we cannot hope for a finite number of distinct results. We now define a small set $\overline{\Theta}$ of pairs of points with the following weaker property: for any points $p, q \in \mathbb{R}^2$, there are two points $p', q' \in \overline{\Theta}$ such that $|\pi(p, q) - \pi(p', q')| \le \varepsilon/2$. We now choose $s = s(\varepsilon/2)$ in Eq. (2) so that $|\hat{\pi}(p, q) - \pi(p, q)| \le \varepsilon/2$ for all pairs $p, q \in \overline{\Theta}$ with probability at least $1 - \delta$. Then the above property of $\overline{\Theta}$ implies that $|\hat{\pi}(p, q) - \pi(p, q)| \le \varepsilon/2$ for all $p, q \in \mathbb{R}^2$ with probability at least $1 - \delta$.

We show that the continuous distribution of each vertex v can be approximated by a discrete distribution of size $O((n^2/\varepsilon^2)\log(n/\delta))$, and then reduce the problem to the discrete case.

For parameters $\alpha > 0$ and $\delta' \in (0, 1)$, set $\nu(\alpha) = \frac{c}{\alpha^2}\log\frac{1}{\delta'}$, where c is a constant. For each $v_i \in V$, we choose a random sample $\overline{\mathsf{H}}_i$ of size $\nu(\alpha)$, according to the pdf γ_i. We regard $\overline{\mathsf{H}}_i$ as a uniform discrete distribution. Let $\overline{\mathsf{H}} = \overline{\mathsf{H}}_1 \times \ldots \times \overline{\mathsf{H}}_n$ be the resulting discrete distribution of the height function.

Consider the case when p, q are fixed, the heights of $v_1, \ldots, v_{n-1}$ are fixed at, say, $x_1, \ldots, x_{n-1}$, and $h(v_n)$ is drawn from a continuous distribution defined by the pdf γ_n. Set

$$J_n(x_1, \ldots, x_{n-1}) = \{x \in \mathbb{R} \mid p \sim_h q \wedge h = \langle x_1, \ldots, x_{n-1}, x\rangle\}.$$

Lemma 3.5. *For any $x_1, \ldots, x_{n-1} \in \mathbb{R}$, $J_n(x_1, \ldots, x_{n-1})$ consists of at most two connected components.*

Proof. Fix $x_1, \ldots, x_{n-1}$ and let $J_n = J_n(x_1, \ldots, x_{n-1})$. If $J_n = \emptyset$, then the lemma is obviously true, so assume that $J_n \neq \emptyset$. Let I be a connected component of J_n, and let $x \in I$ be a point. Suppose $\rho_{\overline{h}}(p)$ and $\rho_{\overline{h}}(q)$, where $\overline{h} = \langle x_1, \ldots, x_{n-1}, x\rangle$, lie on an edge (u, t) of $\mathsf{T}_{\overline{h}}$. There are two cases:

(i) $\rho_{\overline{h}}(v_n)$ has no chance of lying in between $\rho_{\overline{h}}(p)$ and $\rho_{\overline{h}}(q)$ on $\mathsf{T}_{\overline{h}}$. In this case, p, q will always lie on the edge (u, t), and $I = J_n$;

(ii) $\rho_{\overline{h}}(v_n)$ has a chance lying in between $\rho_{\overline{h}}(p)$ and $\rho_{\overline{h}}(q)$ on $\mathsf{T}_{\overline{h}}$. In this case, as x varies, $\rho_{\overline{h}}(p)$, $\rho_{\overline{h}}(q)$ first lie on the edge (u, t), then $\rho_{\overline{h}}(p)$ (resp. $\rho_{\overline{h}}(q)$) lies on the edge (u, v) (resp. (t, v)), then again $\rho_{\overline{h}}(p)$, $\rho_{\overline{h}}(q)$ lie on the edge (u, t). Hence J_n has two connected components.

This concludes the proof. $\square$

For a height function $h = \langle x_1, \ldots, x_n\rangle$, let

$$\mathbf{1}(p, q; x_1, \ldots, x_n) = \mathbf{1}(p \sim_h q).$$

For fixed $x_1, \ldots, x_{n-1}$, let $\pi(p, q \mid x_1, \ldots, x_{n-1})$ denote the conditional probability of $p \sim_h q$ provided that $h(v_j) = x_j$ for all $j < n$. That is

$$\begin{aligned}\pi(p, q \mid x_1, \ldots, x_{n-1}) &= \int_{\mathbb{R}} \gamma_n(x) \cdot \mathbf{1}(p, q; x_1, \ldots, x_{n-1}, x)\mathrm{d}x \\ &= \int_{J_n(x_1, \ldots, x_{n-1})} \gamma_n(x)\mathrm{d}x.\end{aligned}$$

Similarly, we define $\overline{\pi}(p, q \mid x_1, \ldots, x_{n-1})$ as the conditional probability of $p \sim_h q$ provided that $h(v_j) = x_j$ for $j < n$, and $h(v_n)$ is drawn from the uniform discrete distribution $\overline{\mathsf{H}}_n$. That is

$$\begin{aligned}\overline{\pi}(p, q \mid x_1, \ldots, x_{n-1}) &= \frac{1}{\nu(\alpha)} \sum_{j=1}^{\nu(\alpha)} \mathbf{1}(p, q; x_1, \ldots, x_{n-1}, x_n^j) \\ &= \frac{1}{\nu(\alpha)} |J_n(x_1, \ldots, x_{n-1}) \cap \overline{\mathsf{H}}_n|.\end{aligned}$$

Since $J_n(x_1, \ldots, x_{n-1})$ can be represented as at most two connected intervals, a classic result on random sampling by Vapnik and Chervonenkis [24] (see also [18]) implies that

$$|\pi(p, q \mid x_1, \ldots, x_{n-1}) - \overline{\pi}(p, q \mid x_1, \ldots, x_{n-1})| \le \alpha, \quad (3)$$

with probability at least $1 - \delta'$, provided that the constant in $\nu(\alpha)$ is chosen sufficiently large, i.e., $\overline{\mathsf{H}}_n$ approximates γ_n.

We now let $\overline{\pi}(p, q)$ denote the probability of $p \sim_h q$ if h is drawn from the discrete distribution $\overline{\mathsf{H}}$;

$$\overline{\pi}(p, q) = \frac{1}{\nu^n(\alpha)} \sum_{\overline{h} \in \overline{\mathsf{H}}} \mathbf{1}(p \sim_{\overline{h}} q). \quad (4)$$

Lemma 3.6. *For any $p, q \in \mathbb{R}^2$, $|\pi(p, q) - \overline{\pi}(p, q)| \le \alpha n$, with probability at least $1 - \delta'$.*

Proof. Recall that

$$\pi(p, q) = \int \cdots \int \prod_{i=1}^{n} \gamma_i(x_i) \cdot \mathbf{1}(p, q; x_1, \ldots, x_n)\mathrm{d}x_n \cdots \mathrm{d}x_1.$$

Since (3) holds for all pairs $p, q \in \mathbb{R}^2$ and for all $x_1, \ldots, x_{n-1} \in \mathbb{R}$, we obtain

$$\pi(p, q) \le \frac{1}{\nu(\alpha)} \sum_{j_n=1}^{\nu(\alpha)} \int \cdots \int \prod_{i=1}^{n-1} \gamma_i(x_1)$$

$$\cdot\, \mathbf{1}(p,q;x_1,\ldots,x_{n-1},x_n^{j_n})\mathrm{d}x_{n-1}\cdots\mathrm{d}x_1+\alpha.$$

Repeating this step $n-1$ more times and using (4), we obtain

$$\begin{aligned}\pi(p,q) &\le \frac{1}{\nu^n(\alpha)}\sum_{j_1=1}^{\nu(\alpha)}\cdots\sum_{j_n=1}^{\nu(\alpha)}\mathbf{1}(p,q;x_1^{j_1},\ldots,x_n^{j_n})+n\alpha\\ &=\overline{\pi}(p,q)+n\alpha.\end{aligned}$$

Similarly we can prove that $\pi(p,q)\ge\overline{\pi}(p,q)-n\alpha$. □

Setting $\alpha=\varepsilon/2n$, we obtain that $|\pi(p,q)-\overline{\pi}(p,q)|\le\varepsilon/2$ for any pair $p,q\in\mathbb{R}^2$. By choosing $\delta'=\delta/2$, the above inequality holds with probability at least $1-\delta/2$.

Since $\overline{\mathsf{H}}$ is a discrete distribution, invoking Theorem 3.4 for $\overline{\mathsf{H}}$ and using Lemma 3.6, we obtain the following.

Theorem 3.7. *Let $\mathbb{M}=(V,E,F)$ be a triangulation of $\mathbb{R}^2$, where $n=|V|$, let the height of each vertex be described by a continuous distribution such that a random instantiation can be performed in constant time, and let $\varepsilon,\delta\in(0,1)$ be two parameters. A data structure of size $O((n/\varepsilon^2)\log(n/(\varepsilon\delta)))$ can be constructed in $O((n/\varepsilon^2)\log(n/(\varepsilon\delta))\log n)$ time that computes, for any $p,q\in\mathbb{R}^2$, in $O((1/\varepsilon^2)\log(n/(\varepsilon\delta))\log n)$ time, a value $\hat{\pi}(p,q)$ such that $|\hat{\pi}(p,q)-\pi(p,q)|\le\varepsilon$ with probability at least $1-\delta$.*

Remarks. Theorem 3.7 also holds for merge and split trees.

4. THE DISTANCE STATISTICS OF TWO POINTS

Recall the distance function $\mathrm{d}_h(\cdot,\cdot)$ based on a height function, introduced in Section 2. Given a triangulation $\mathbb{M}$ and a distribution H on the height function, we build a data structure to estimate certain statistics on $\mathrm{d}_h(p,q)$ for two query points $p,q\in\mathbb{R}^2$ where h is the random function drawn from H. For simplicity, we assume H to be a discrete distribution of description complexity k. We are interested in the following two statistics: (i) the expected value of $\mathrm{d}_h(p,q)$, denoted by $\overline{\mathrm{d}}(p,q)$, i.e.,

$$\overline{\mathrm{d}}(p,q)=\sum_{h\in\mathsf{H}}\gamma(h)\mathrm{d}_h(p,q),$$

and (ii) given $\ell>0$, the probability of $\mathrm{d}_h(p,q)$ being at least ℓ, denoted by $\varphi(p,q;\ell)$, i.e.,

$$\varphi(p,q;\ell)=\Pr[\mathrm{d}_h(p,q)\ge\ell]=\sum_{h:\mathrm{d}_h(p,q)\ge\ell}\gamma(h).$$

As in Section 3, we use a simple Monte Carlo algorithm for estimating $\overline{\mathrm{d}}(p,q)$ and $\varphi(p,q;\ell)$. Namely, we fix a parameter $s\ge1$. For each $i\le s$, we choose a random height function $h_i\in\mathsf{H}$ and construct in $O(n\log n)$ time a linear-size data structure so that for any pair $p,q\in\mathbb{R}^2$, $\mathrm{d}_{h_i}(p,q)$ can be computed in $O(\log n)$ time. For a query pair $p,q\in\mathbb{R}^2$, we compute $\mathrm{d}_{h_i}(p,q)$ for all $i\le s$. We return $\widehat{\mathrm{d}}(p,q)=\frac{1}{s}\sum_{i=1}^{s}\mathrm{d}_{h_i}(p,q)$ as an estimate for $\overline{\mathrm{d}}(p,q)$ and $\widehat{\varphi}(p,q;\ell)=|\{i\mid\mathrm{d}_{h_i}(p,q)\ge\ell\}|/s$ as an estimate of $\varphi(p,q;\ell)$.

The query time, size and preprocessing time are $O(s\log n)$, $O(sn)$ and $O(nk+sn\log n)$, respectively. In the rest of the section we obtain bounds on s to ensure a good estimation of $\overline{\mathrm{d}}(\cdot,\cdot)$ and $\varphi(\cdot,\cdot;\cdot)$.

Analysis for expected distance. We begin by introducing a few definitions. For a vertex v_i, let $h_i^+=\max_{1\le j\le k}h_i^j$, $h_i^-=\min_{1\le j\le k}h_i^j$, $\Delta_i=h_i^+-h_i^-$; set $\Delta=\max_{1\le i\le n}\Delta_i$. For a path χ in $\mathbb{R}^2$ and for a height function $h\in\mathsf{H}$, let $\|\chi\|_h=\max_{x\in\chi}h(x)-\min_{x\in\chi}h(x)$. For a pair of points $p,q\in\mathbb{R}^2$ and for a height function $h\in\mathsf{H}$, let $\psi_h(p,q)$ denote a path in $\mathbb{R}^2$ such that $\|\psi_h(p,q)\|_h=\mathrm{d}_h(p,q)$; i.e., $\psi_h(p,q)$ is a minimum height-difference path on Σ_h.

Lemma 4.1. *For any pair $p,q\in\mathbb{R}^2$, there exists a value $\lambda_{p,q}$ such that for any height function $h\in\mathsf{H}$, $\mathrm{d}_h(p,q)\in[\lambda_{p,q}-\Delta,\lambda_{p,q}+\Delta]$.*

Proof. Consider the height function $h^-=\langle h_1^-,\ldots,h_n^-\rangle$. Let $\psi^-=\psi_{h^-}(p,q)$ and $\lambda_{p,q}=\mathrm{d}_{h^-}(p,q)$.

For any $h\in\mathsf{H}$ and for any $i\le n$, $h(v_i)\in[h_i^-,h_i^-+\Delta]$, therefore,

$$\mathrm{d}_h(p,q)\le\|\psi^-\|_h\le\|\psi^-\|_{h^-}+\Delta=\lambda_{p,q}+\Delta.$$

Similarly, we can argue that for any $h\in\mathsf{H}$, $\lambda_{p,q}\le\mathrm{d}_h(p,q)+\Delta$. These two inequalities together imply the lemma. □

The following well-known lemma (see e.g. [9]) gives a tail estimate on function values, with bounded range, over random variables.

Lemma 4.2. *(Hoeffding) Let $x_1,\ldots,x_s$ be s i.i.d. random variables with $f(x)\in[a,b]$. Then for all $\varepsilon>0$,*

$$\Pr\Big[\Big|\frac{1}{s}\sum_{i=1}^{s}f(x_i)-\mathsf{E}[f(x)]\Big|>\varepsilon\Big]\le2\exp\Big(-\frac{2s\varepsilon^2}{(b-a)^2}\Big).$$

For a pair $p,q\in\mathbb{R}^2$, let $\mathsf{err}(p,q)=|\widehat{\mathrm{d}}(p,q)-\overline{\mathrm{d}}(p,q)|$. Then for a fixed pair $p,q\in\mathbb{R}^2$, Lemmas 4.1 and 4.2 imply

$$\Pr[\mathsf{err}(p,q)>\varepsilon\Delta]\le2\exp\Big(-\frac{s\varepsilon^2}{2}\Big). \qquad (5)$$

To bound $\mathsf{err}(p,q)$, we follow an argument similar to Section 3. We construct the overlay of extended height level maps of all height functions $h\in\mathsf{H}$, and we also overlay $\mathbb{M}$ on it. Finally, we triangulate each face of the overlay. Let Ξ denote the resulting planar subdivision — each cell of Ξ lies in a single triangle of $\mathbb{M}$ as well as in a single face of all extended height-level maps. Let Ω denote the set of vertices of Ξ, and let $\Theta=\Omega\times\Omega$. $|\Omega|=O(n^3k^8)$ and $|\Theta|=O(n^6k^{16})$.

The following lemma states the desired property of Ξ.

Lemma 4.3. *Let τ_1,τ_2 be two triangles in Ξ, and let $p_1,q_1\in\tau_1$, $p_2,q_2\in\tau_2$. For all $h\in\mathsf{H}$, the following conditions are satisfied:*

(i) $p_1\sim_h q_1$, and $p_2\sim_h q_2$.

(ii) If the maximum-height point on $\psi_h(p_1,p_2)$ is a vertex v of $\mathbb{M}$ (resp. an endpoint p_1,p_2), then the maximum-height point on $\psi_h(q_1,q_2)$ is also v (resp. q_1,q_2).

(iii) If the minimum-height point on $\psi_h(p_1,p_2)$ is a vertex w of $\mathbb{M}$ (resp. an endpoint p_1,p_2), then the minimum-height point on $\psi_h(q_1,q_2)$ is also w (resp. q_1,q_2).

Proof. (i) follows from the construction. We prove (ii); (iii) is symmetric. Suppose the maximum-height endpoint of $\psi_h(p_1,p_2)$ is a vertex v, but the maximum-height endpoint of $\psi_h(q_1,q_2)$ is q_1. Then $h(p_1)<h(v)<h(q_1)$, but then the level set of v w.r.t. h separates p_1 and q_1. This contradicts the assumption that p_1,q_1 lie in the same face of Ξ. □

Let x, y be two points in $\mathbb{R}^2$, and let τ_x, τ_y be the triangles containing x and y, respectively. Using Lemma 4.3 and the fact for any $h \in \mathsf{H}$, $h(x)$ (resp. $h(y)$) can be written as a convex combination of the heights of the vertices of τ_x (resp. τ_y), we can prove the following lemma. The proof is rather tedious, and omitted in this paper.

Lemma 4.4. *If* $\mathsf{err}(p, q) \le \varepsilon\Delta$ *for every* $(p, q) \in \Theta$*, then* $\mathsf{err}(p, q) \le 3\varepsilon\Delta$ *for every* $(p, q) \in \mathbb{R}^2 \times \mathbb{R}^2$*.*

Setting $s = O(\frac{1}{\varepsilon^2} \log \frac{|\Theta|}{\delta})$, we obtain the following.

Theorem 4.5. *Let* $\mathbb{M}$ *be a triangulation of* $\mathbb{R}^2$*, let* H *be a discrete distribution on the height function of description complexity* k*, and let* $\varepsilon, \delta \in (0, 1)$ *be two parameters. Let* $s = O(\frac{1}{\varepsilon^2} \log \frac{nk}{\delta})$*. A data structure of size* $O(sn)$ *can be constructed in* $O(sn \log n + nk)$ *time that for any* $p, q \in \mathbb{R}^2$*, computes* $\overline{\mathrm{d}}(p, q)$ *within additive error* $\varepsilon\Delta$ *with probability at least* $1 - \delta$*. Here* Δ *is the maximum variation in the height of a vertex of* $\mathbb{M}$ *in* H*.*

Remarks. The analysis can be extended to continuous distributions using the same idea as in Section 3, i.e., we can show that the continuous distribution can be approximated by a discrete distribution.

Analysis for tail probability. Next we bound s, the number of samples, required for estimating $\varphi(p, q; \ell)$ within additive error ε. For a fixed triple $(p_0, q_0, \ell_0) \in \mathbb{R}^5$, where $p_0, q_0 \in \mathbb{R}^2$ and $\ell_0 \in \mathbb{R}$, the Chernoff bound, as in Section 3, implies that

$$\Pr[|\varphi(p_0, q_0; \ell_0) - \widehat{\varphi}(p_0, q_0; \ell_0)| \ge \varepsilon] \le 2\exp(-2\varepsilon^2/s). \quad (6)$$

As earlier, we construct a representative set Θ, but now $\Theta \subset \mathbb{R}^5$, so that if (6) holds for all triples in Θ, it also holds for all triples in $\mathbb{R}^5$.

We begin by constructing Ξ, the overlay of all possible extended height-level maps, as above. For a cell $\zeta \in \Xi$, let H_ζ denote the set of all possible height functions for ζ. Since ζ lies inside a triangle of $\mathbb{M}$, H_ζ is a set of k^3 linear functions. Fix a pair ζ, η of cells of Ξ. For $1 \le i, j \le k^3$, define a 4-variate linear function $g^{ij}_{\zeta\eta} : \mathbb{R}^4 \to \mathbb{R}$ as follows:

$$g^{ij}_{\zeta\eta}(p, q) = h^i_\zeta(p) - h^j_\eta(q),$$

where $p, q \in \mathbb{R}^2$ and $h^i_\zeta \in \mathsf{H}_\zeta$, $h^j_\eta \in \mathsf{H}_\eta$. For each possible height h^r_l of vertex v_l, where $1 \le l \le n, 1 \le r \le k$, we define

$$g^{ir}_{\zeta l}(p, q) = h^i_\zeta(p) - h^r_l \text{ and } g^{jr}_{\eta l}(p, q) = h^j_\eta(q) - h^r_l.$$

Set

$$G_{\zeta\eta} = \{g^{ij}_{\zeta\eta}, -g^{ij}_{\zeta\eta}, g^{ir}_{\zeta l}, -g^{ir}_{\zeta l}, g^{jr}_{\eta l}, -g^{jr}_{\eta l} \mid i, j \in [k^3], l \in [n], r \in [k]\}.$$

The graph of each linear function in $G_{\zeta\eta}$ is a hyperplane in $\mathbb{R}^5$, so we will also regard $G_{\zeta\eta}$ as a set of hyperplanes in $\mathbb{R}^5$. $|G_{\zeta\eta}| = O(nk^4 + k^6)$.

The *arrangement* of $G_{\zeta\eta}$, denoted by $\mathcal{A}(G_{\zeta\eta})$, is the decomposition of $\mathbb{R}^5$ into maximal connected regions each of which lies in the same subset of hyperplanes of $G_{\zeta\eta}$. We clip $\mathcal{A}(G_{\zeta\eta})$ within $\zeta \times \eta \times \mathbb{R} = \{(p, q, \ell) \mid p \in \zeta, q \in \eta, \ell \in \mathbb{R}\}$. It is known that each cell of $\mathcal{A}(G_{\zeta\eta})$ is convex, and $\mathcal{A}(G_{\zeta\eta})$ has $O(|G_{\zeta\eta}|^5)$ cells (see the survey [3] for details on arrangements). We choose a point $(p_\tau, q_\tau, \ell_\tau)$ from each cell τ of $\mathcal{A}(G_{\zeta\eta})$. We repeat this process for all pairs $\zeta, \eta \in \Xi$, and let Θ denote the resulting set of triples; $|\Theta| = O((nk)^{O(1)})$.

Lemma 4.6. *For any triple* $(p, q, \ell) \in \mathbb{R}^5$*, there is a triple* $(p_\tau, q_\tau, \ell_\tau) \in \Theta$ *such that* $\varphi(p, q; \ell) = \varphi(p_\tau, q_\tau; \ell_\tau)$*.*

Proof. For a triple $(p, q, \ell) \in \mathbb{R}^5$, let $\mathsf{H}(p, q; \ell) = \{h \in \mathsf{H} \mid \mathrm{d}_h(p, q) \ge \ell\}$. Fix a triple $(p_0, q_0, \ell_0) \in \mathbb{R}^5$. Let ζ (resp. η) be the cell of Ξ that contains p_0 (resp. q_0), and let τ be the cell of $\mathcal{A}(G_{\zeta\eta})$ that contains (p_0, q_0, ℓ_0). We claim that $\mathsf{H}(p_0, q_0; \ell_0) = \mathsf{H}(p_\tau, q_\tau; \ell_\tau)$, which would imply the lemma.

Suppose to the contrary. Let $h \in \mathsf{H}$ be a height function such that $h \in \mathsf{H}(p_0, q_0; \ell_0) \oplus \mathsf{H}(p_\tau, q_\tau; \ell_\tau)$. Without loss of generality, assume that $\mathrm{d}_h(p_0, q_0) \ge \ell_0$, and $\mathrm{d}_h(p_\tau, q_\tau) < \ell_\tau$. Since $p_0, p_\tau \in \zeta$ and $q_0, q_\tau \in \eta$, (p_0, q_0) and (p_τ, q_τ) satisfy Lemma 4.3. Consequently both the highest and the lowest points of $\psi_h(p_0, q_0)$ cannot be vertices of $\mathbb{M}$ because then $\mathrm{d}_h(p_\tau, q_\tau) = \mathrm{d}_h(p_0, q_0)$. Hence, at least one of them is an endpoint of $\psi_h(p_0, q_0)$.

For simplicity, assume both extremal points of $\psi_h(p_0, q_0)$ are its endpoints and $h(p_0) \ge h(q_0)$; the argument for the other cases is similar. Then $\mathrm{d}_h(p, q) = |h(p) - h(q)|$ for all $(p, q) \in \zeta \times \eta$. In other words, $\mathrm{d}_h(p, q) = |h^i_\zeta(p) - h^j_\eta(q)|$ for some $h^i_\zeta \in \mathsf{H}_\zeta$ and $h^j_\eta \in \mathsf{H}_\eta$. That is, $\mathrm{d}_h(p_0, q_0) = h^i_\zeta(p_0) - h^j_\eta(q_0) = g^{ij}_{\zeta\eta}(p_0, q_0) \ge \ell_0$. On the other hand, $\mathrm{d}_h(p_\tau, q_\tau) < \ell$ implies that $g^{ij}_{\zeta\eta}(p_0, q_0) < \ell_0$. In other words, the line segment connecting (p_0, q_0, ℓ_0) and $(p_\tau, q_\tau, \ell_\tau)$ intersects the hyperplane $g^{ij}_{\zeta\eta}$. Since each cell of $\mathcal{A}(G_{\zeta\eta})$ is convex, the segment intersecting $g^{ij}_{\zeta\eta}$ implies that $(p_0, q_0, \ell_0) \notin \tau$, which contradicts with the assumption that $(p_0, q_0, \ell_0) \in \tau$. Hence $\mathsf{H}(p_0, q_0; \ell_0) = \mathsf{H}(p_\tau, q_\tau; \ell_\tau)$, as desired. $\Box$

Setting $s = O(\frac{1}{\varepsilon^2} \log \frac{|\Theta|}{\delta}) = O(\frac{1}{\varepsilon^2} \log \frac{nk}{\delta})$, we obtain the following.

Theorem 4.7. *Let* $\mathbb{M}$ *be a triangulation of* $\mathbb{R}^2$*, let* H *be a discrete distribution on the height function of description complexity* k*, and let* $\varepsilon, \delta \in (0, 1)$ *be two parameters. Let* $s = O(\frac{1}{\varepsilon^2} \log \frac{nk}{\delta})$*. A data structure of size* $O(sn)$ *can be constructed in* $O(sn \log n + nk)$ *time that for any* $p, q \in \mathbb{R}^2$ *and* $\ell \in \mathbb{R}$*, computes* $\varphi(p, q; \ell)$ *within additive error* ε *with probability at least* $1 - \delta$*.*

5. APPLICATIONS

In this section we briefly describe two applications of the algorithms in Sections 3 and 4, both motivated by hydrology analysis of terrains. Due to lack of space, we omit some details from here.

Topological persistence was introduced by Edelsbrunner et al. [14] and can roughly be defined as follows. Suppose we sweep a horizontal plane in the direction of increasing values of h and keep track of connected components in $\mathbb{M}_{<\ell}$ (sometimes referred to as *basins*) while increasing ℓ. A component of $\mathbb{M}_{<\ell}$ starts at a minimum vertex and ends at a saddle vertex when it joins with another component. Suppose two components of C_1, C_2 of $\mathbb{M}_{<\ell}$ with minima α_1 and α_2 respectively merge at a saddle β with $h(\beta) = \ell$. Assume $h(\alpha_1) > h(\alpha_2)$. Then C_1 "ends" at β and the merged component of $\mathbb{M}_{<\ell}$ becomes C_2. The persistence of β, denoted by $\omega(\beta)$, is $\omega(\beta) = h(\beta) - h(\alpha_1)$. We refer to C_1 as the basin associated with β and denote it by B_β. We say that $\omega(\beta)$ is the persistence of the basin B_β. We also set $\omega(\alpha_1) = \omega(\beta)$. A common operation in topological simplification and hydrology analysis is to simplify, fill, or remove the basins whose persistence is below a given threshold θ [4, 12, 13].

A point $p \in \mathbb{R}^2$ lies in a sequence of nested basins. Among them the one with the lowest persistence is called the *elementary basin* of p and is defined by B_p. Note that if p lies on the edge (α, β) of the merge tree, with $h(\alpha) < h(\beta)$, then $B_p = B_\beta$. A useful query is to ask for the persistence of B_p for a query point $p \in \mathbb{R}^2$. In the context of uncertain terrains, given $p \in \mathbb{R}^2$ and $\theta > 0$, we wish to compute $\Pr[\omega(B_p) \geq \theta]$.

The probability is the same for any two points that lie in the same cell of the overlay of the height-level maps, denoted by $\widehat{\mathsf{M}}$ in Section 3, so the analysis of Section 3 can be adapted for this case as well. In particular, by choosing a set of $s = O(\frac{1}{\varepsilon^2} \log \frac{nk}{\delta})$ samples, we can estimate $\Pr[\omega(B_p) \geq \theta]$ within error ε with probability at least $1 - \delta$.

Theorem 5.1. *Let $\mathbb{M}$ be a triangulation of $\mathbb{R}^2$, let H be a discrete distribution on the height function of description complexity k, and let $\varepsilon, \delta \in (0, 1)$ be two parameters. Let $s = O(\frac{1}{\varepsilon^2} \log \frac{nk}{\delta})$. A data structure of size $O(sn)$ can be constructed in $O(sn \log n + nk)$ time that for any query point $p \in \mathbb{R}^2$ and for any $\theta > 0$ can estimate $\Pr[\omega(B_p) \geq \theta]$ within additive error ε in time $O(s \log n)$, with probability at least $1 - \delta$.*

Another related query that arises in hydrology analysis is as follows. Let $\mathcal{M}_h$ denote the merge tree of the height function h, and let $\rho : \mathbb{M} \to \mathcal{M}_h$ be the retraction map from $\mathbb{M}$ to $\mathcal{M}_h$, analogous to the one described for contour trees. For two points $p, q \in \mathbb{R}^2$, let $\lambda(p, q)$ denote the nearest common ancestor of $\rho(p)$ and $\rho(q)$ in $\mathcal{M}_h$. We define $\sigma(p, q) = h(\lambda(p, q)) - h(p)$; we will be interested in the case when p is a minimum. Then $\sigma(p, q)$ intuitively tells us how much the water level has to rise at a "pit" p before it spills to q (see e.g. [4]). Given p, q, we wish to compute $\sigma(p, q)$. In the context of uncertain terrains, we want to either compute the expected value of $\sigma(p, q)$, or given a value ℓ, estimate $\Pr[\sigma(p, q) \geq \ell]$.

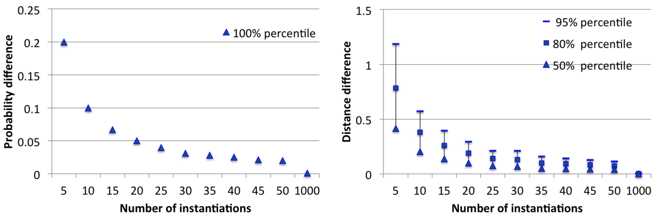
The algorithm and the analysis described in Section 4 works for this problem. Hence by choosing $s = O(\frac{1}{\varepsilon^2} \log \frac{nk}{\delta})$, we can compute the expected value within error $\varepsilon\Delta$ and the probability $\Pr[\sigma(p, q) \geq \ell]$ within error ε, with probability at least $1 - \delta$.

Theorem 5.2. *Let $\mathbb{M}$ be a triangulation of $\mathbb{R}^2$, let H be a discrete distribution on the height function of description complexity k, and let $\varepsilon, \delta \in (0, 1)$ be two parameters. Let $s = O(\frac{1}{\varepsilon^2} \log \frac{nk}{\delta})$. A data structure of size $O(sn)$ can be constructed in $O(sn \log n + nk)$ time that for any $p, q \in \mathbb{R}^2$ and $\ell \in \mathbb{R}$, computes in $O(s \log n)$ time the expected value of $\sigma(p, q)$ within additive error $\varepsilon\Delta$ as well as $\Pr[\sigma(p, q) \geq \ell]$ within additive error ε, with probability at least $1 - \delta$. Here Δ is the maximum variation in the height of a vertex of $\mathbb{M}$ in H.*

6. EXPERIMENTS

We have conducted experiments on a real dataset to demonstrate the efficacy of our methods for estimating $\pi(p, q)$, the probability of p, q lying on an edge of the contour tree, and for estimating $\overline{\mathrm{d}}(p, q)$, the expected distance of p, q on the contour tree.

Datasets. We use the terrain dataset San Bernardino, which is a grid composed of $128 \times 128 = 16384$ vertices. We use the upper-left quarter of the data[1], and triangulate it, resulting in $n = 4096$ vertices, $m = 12033$ edges, and $t = 7938$ triangles in its underlying triangulation. Its (exact) vertex heights are scaled and translated into the range $[0, 1000]$, for the ease of later discussions. We also tested using two smaller synthetic datasets, one shown in Fig. 3(a), and the other based on a grid of size $17 \times 17 = 289$ vertices with randomly-generated (exact) heights. The results are very similar to each other, so we focus on the San Bernardino dataset. All these datasets do not have uncertainty on the vertex heights, and we introduce uncertainty below.



(a) (b)

Figure 4. The difference between empirical and exact values of $\pi(p, q)$, and the difference between empirical and exact values of $\overline{\mathrm{d}}(p, q)$.

Uncertainty. Given a parameter Δ, we introduce uncertainty on the vertex heights as follows. For any vertex v, let h_v denote its height in the above dataset (without uncertainty). We consider both continuous and discrete distributions to model the vertex heights. The continuous distributions we consider are a uniform distribution $\mathrm{U}(h_v - \Delta/2, h_v + \Delta/2)$, a Gaussian distribution $\mathrm{N}(h_v, \Delta/6)$, and a mixture of two Gaussians specified by $\mathrm{N}(h_v - \Delta/4, \Delta/12)$ and $\mathrm{N}(h_v + \Delta/4, \Delta/12)$ with equal mixing probabilities. The discrete distributions were over k possible heights where the values for the heights were generated by drawing k numbers from the uniform distribution $\mathrm{U}(h_v - \Delta/2, h_v + \Delta/2)$. Given the k values we specified two models to generate uncertain data. The first model was a single draw from the uniform distribution over the k values. The second model was a single draw from a multinomial distribution over the k values with a random probability vector, where the probability vector was a random draw from the k-simplex. In our experiments, we set $k = 5$ and $\Delta \in \{10, 20, 30, 40, 50\}$.

Queries. Among all $t = 7938$ triangles in the underlying triangulation of the dataset, we randomly select 100 of them. For each selected triangle, we choose its centroid as the representative point. Each two distinct representative points p, q constitute one query (p, q), resulting in $\binom{100}{2} = 4950$ number of queries.

Obtaining the exact values of $\pi(p, q)$ and $\overline{\mathrm{d}}(p, q)$. Computing $\pi(p, q)$ and $\overline{\mathrm{d}}(p, q)$ exactly even for these moderate size datasets is exorbitantly expensive. Instead, we apply our Monte-Carlo methods for a sufficient number of times, and use the resulting estimates as the exact values. In our experiments, we found that 1,000 samples are sufficient enough (this will be verified later). We did not present results on larger datasets because it is too expensive to generate 1,000 samples.

Measuring the convergence. For each query (p, q) and

[1] We do this simply to constrain memory requirements.

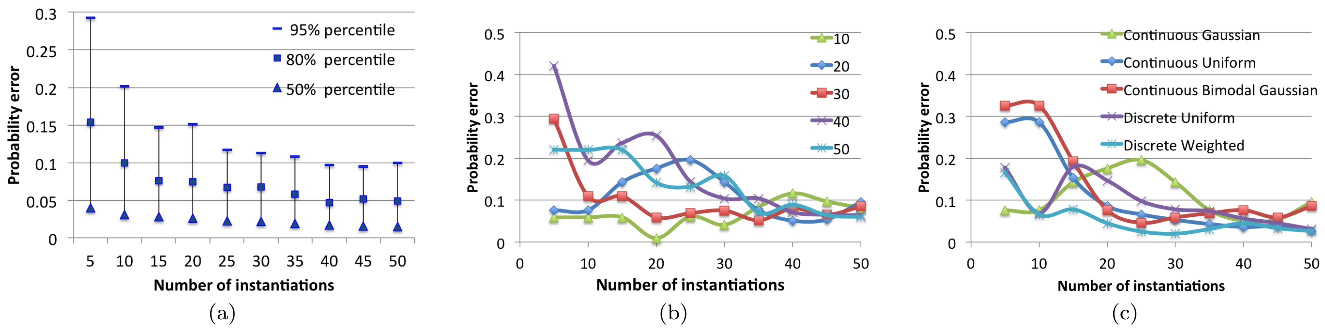


Figure 5. Accuracy of $\pi(p,q)$.

round j, we compute the difference $|\hat{\pi}_j(p,q) - \hat{\pi}_{j-1}(p,q)|$, where $\hat{\pi}_j(p,q)$ denotes the estimate of $\pi(p,q)$ in the j-th round. Among all $\binom{100}{2} = 4950$ queries, we report the maximum (i.e., the 100-th percentile) of these probability differences. Analogously, we consider the distance differences, and we report the 50-th, 80-th and 95-th percentiles of these distance differences.

Measuring the effectiveness. For each query (p,q), we compute the errors in probability estimates $|\pi(p,q) - \hat{\pi}(p,q)|$ and distance estimates $\left|\overline{\mathrm{d}}(p,q) - \widehat{\mathrm{d}}(p,q)\right|$. Among all $\binom{100}{2} = 4950$ queries, we report the 50-th, 80-th and 95-th percentiles of these errors. Note that if $\pi(p,q) = 0$, then $\hat{\pi}(p,q) = 0$ and the error is 0; the error is also 0 when $\pi(p,q) = 1$. We discard such queries when we measure the errors; including them will only reduce the error.

Convergence of our methods. We tested how our methods converged as we varied s, the number of instantiations. Fig. 4(a)–(b) illustrate that the differences in both probability and distance decrease very quickly, and when $s = 1000$ the differences are effectively zero. Therefore, we take the empirical estimates with 1000 samples as the exact values. For Fig. 4(a)–(b), we used a Gaussian distribution and $\Delta = 20$.

Estimating $\pi(p,q)$. We examined the accuracy of our estimate of $\pi(p,q)$ as we varied s, the number of instantiations, from 5 to 50. See Fig. 5(a)–(c). Not surprisingly, as s increases our estimates improve and errors are reasonably small when $s \geq 20$. The smaller uncertainty (as denoted by Δ) also results in more accurate estimates, though Fig. 5(b) does not convey a very clear pattern. Regarding the various generative distributions of uncertainty, the discrete multinomial distribution provided the best estimates. For Fig. 5(a) we used a Gaussian distribution with $\Delta = 20$. For Fig. 5(b)-(c), we used the Gaussian distribution with $\Delta = 20$, and we report the 95-th percentile.

Estimating $\overline{\mathrm{d}}(p,q)$. Fig. 6(a)–(c) show that, as the number of instantiations s increases, the distance estimate is more accurate and the error in the estimate is far smaller than the uncertainty level Δ. In Fig. 6(b) we see that the smaller the uncertainty parameter Δ the more accurate we are in estimating the distance. Fig. 6(c) suggests that the Gaussian distribution and discrete distributions result in more accurate estimates. Fig. 6(a)–(c) are generated using the same simulation parameters as those for Fig. 5(a)–(c).

Distribution of errors vs exact values. We examined the distribution of errors as a function of the exact probability or distance. Fig. 7(a)-(b) illustrate that the estimation errors of probabilities and distances are independent of the underlying exact probability and distance values, respectively. For these simulations we used the data set as shown in Fig. 3(a), Gaussian distribution, $\Delta = 10$, and $s = 20$. The number of queries is $\binom{40}{2} = 780$, as this dataset has only 40 triangles. Note that Fig. 7(a) looks sparser than Fig. 7(b), since there are many $(1,0)$ and $(0,0)$ points in Fig. 7(a).

Distribution of probability estimates. We tested how the percentage of queries with $\pi(p,q) = 0$ or $\pi(p,q) = 1$ or $\pi(p,q) \in (0,1)$ varied as we increased the uncertainty level Δ. Not surprisingly, as we increase the uncertainty level, the percentage of queries with $\pi(p,q) = 1$ decreases, and the percentage of queries with $\pi(p,q) \in (0,1)$ increases, while the percentage of queries with $\pi(p,q) = 0$ decreases very slightly. Fig. 7(c) are based on simulations with the same setting as Fig. 7(a)-(b).

7. CONCLUSION

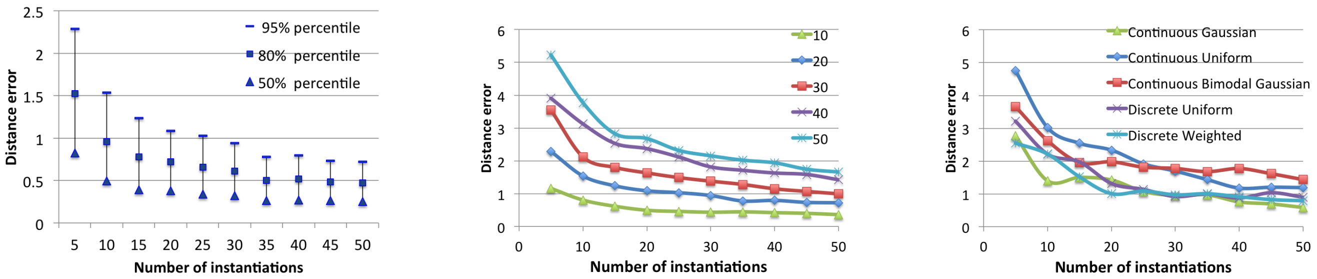
In this paper we studied contour trees of terrains in a probabilistic setting. We presented efficient sampling-based methods for estimating, with high probability, (i) the probability that two points lie on the same edge of the contour tree, within additive error; (ii) the expected distance of two points p,q and the probability that the distance of p,q is at least ℓ on the contour tree, within additive error. We also conducted some preliminary experiments to demonstrate the effectiveness of our methods. We conclude this paper with some open problems: (i) How hard is it to compute the probability of two points lying on an edge of the contour tree *exactly*? What about the distance statistics of two points? (ii) What is a robust and useful contour tree representation of a terrain in the presence of data uncertainty?

Acknowledgements. Work by P.A. and W.Z. is supported by NSF under grants CCF-09-40671, CCF-10-12254, CCF-11-61359, and IIS-14-08846, and by Grant 2012/229 from the U.S.-Israel Binational Science Foundation. S.M. is supported by grants AFOSR: FA9550-10-1-0436, NSF DMS-1418261, NSF DMS-1209155, and NSF IIS-1320357. The authors thank Jungwoo Yang for providing his original contour-tree code and code support for the experiments. The San Bernardino dataset was provided by Paola Magillo and Kenneth Weiss.

8. REFERENCES

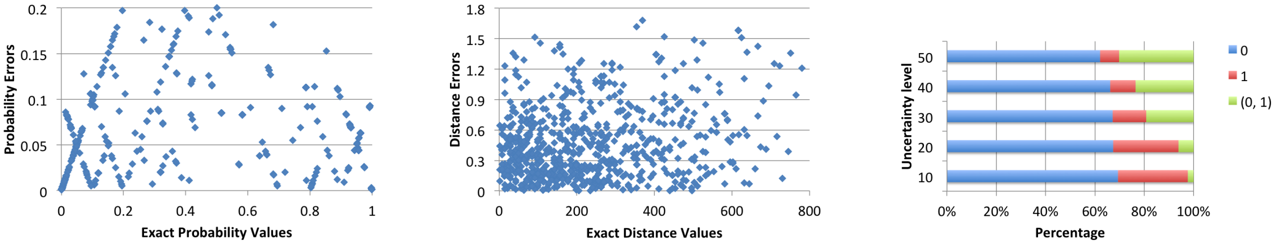
[1] P. K. Agarwal, L. Arge, T. Mølhave, M. Revsbæk, and J. Yang. Maintaining contour trees of dynamic terrains. In *Proc. 31st SoCG*, 796–811, 2015.

[2] P. K. Agarwal, L. Arge, and K. Yi. I/O-efficient batched union-find and its applications to terrain analysis. *ACM Trans. Algs.*, 7:11:1–11:21, 2010.



(a) (b) (c)

Figure 6. Accuracy of $\overline{\mathrm{d}}(p, q)$.



(a) (b) (c)

Figure 7. (a)-(b) Errors vs exact values. (c) The percentage of queries vs uncertainty level.

[3] P. K. Agarwal and M. Sharir. Arrangements and their applications. (J.-R. Sack and J. Urrutia, eds.), *Handbook of Computational Geometry*, 49–119. Elsevier, 2000.

[4] L. Arge, M. Revsbæk, and N. Zeh. I/O-efficient computation of water flow across a terrain. In *Proc. 26th SoCG*, 403–412, 2010.

[5] S. Banerjee, B. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Sptial Data, 2nd ed.* Chapman and Hall, New York, 2015.

[6] U. Bauer, X. Ge, and Y. Wang. Measuring distance between reeb graphs. In *Proc. 30th SoCG*, 464–473, 2014.

[7] K. Bemis, D. Silver, P. Rona, and C. Feng. Case study: a methodology for plume visualization with application to real-time acquisition and navigation. In *Proc. IEEE Vis.*, 481–494, 2000.

[8] M. de Berg and M. J. van Kreveld. Trekking in the alps without freezing or getting tired. In *Proc. 1st ESA*, 121–132, 1993.

[9] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. (O. Bousquet, U. von Luxburg, and G. R atsch, eds.), *Advanced Lectures on Machine Learning*, 169–207. Springer, 2004.

[10] H. Carr, J. Snoeyink, and U. Axen. Computing contour trees in all dimensions. In *Proc. 11th SODA*, 918–926, 2000.

[11] C. Chen, Y. Li, W. Li, and H. Dai. A multiresolution hierarchical classification algorithm for filtering airborne lidar data. *{ISPRS} J. Photo. Remote Sens.*, 82:1–9, 2013.

[12] A. Danner, T. Mølhave, K. Yi, P. K. Agarwal, L. Arge, and H. Mitásová. Terrastream: From elevation data to watershed hierarchies. In *Proc. ACM GIS*, 2007.

[13] H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical morse - smale complexes for piecewise linear 2-manifolds. *Discr. Comput. Geom.*, 30:87–107, 2003.

[14] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *Proc. 41th FOCS*, 454–463, 2000.

[15] C. Gray. Shortest paths on uncertain terrains. MS thesis, Dept. Computer Sci., University of British Columbia, 2004.

[16] C. Gray and W. S. Evans. Optimistic shortest paths on uncertain terrains. In *Proc. 16th CCCG*, 68–71, 2004.

[17] D. Günther, J. Salmon, and J. Tierny. Mandatory critical points of 2D uncertain scalar fields. In *Computer Graphics Forum*, 33, 2014.

[18] S. Har-Peled. *Geometric Approximation Algorithms*. Amer. Math. Soc., 2011.

[19] M. Kraus. Visualization of uncertain contour trees. In *Proc. Int. Conf. Imaging Theory Appl.*, 132–139, 2010.

[20] N. Max, R. Crawfis, and D. Williams. Visualization for climate modeling. *IEEE Comput. Graphics Appl.*, 13:34–40, 1993.

[21] M. Mihai and R. Westermann. Visualizing the stability of critical points in uncertain scalar fields. *Computers & Graphics*, 41:13 – 25, 2014.

[22] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[23] M. van Kreveld, R. van Oostrum, C. Bajaj, V. Pascucci, and D. Schikore. Contour trees and small seed sets for isosurface traversal. In *Proc. 13th SoCG*, 212–220, 1997.

[24] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.

[25] W. Zhang. *Geometric Computing over Uncertain Data*. PhD thesis, Dept. Computer Sci., Duke Univ., 2015.