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 and/or relative 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 also present some experimental results to demonstrate the effectiveness of our approach.

1. INTRODUCTION

In this paper, we study contour trees of terrains in a probabilistic setting. Terrain is generally defined as the vertical and horizontal dimension of land surface, the understanding of which is important in many areas, including but not limited to, geographic information systems, agriculture, hydrology, and aviation. One commonly-studied type of terrain is represented as the graph of a piecewise-linear triangulated surface in $\mathbb{R}^3$, known as *triangulated irregular network* (TIN). This is also the type of our interest in this paper. Due to the inherent measurement errors, it is reasonable to assume that the height of each vertex of the underlying triangulation defining a terrain is described probabilistically.

Contour tree is a fundamental structure for topological analysis and data visualizations on large volume data sets, such as terrains and images. Efficient algorithms have been devised for computing contour trees of terrains in memory [6, 16, 17], I/O-efficiently [2], and for maintaining contour trees of dynamic terrains [1], where terrains are represented as TIN. When the height of each terrain vertex is represented probabilistically, a natural question raises: what is the contour tree of such a terrain? As there can be exponential number of contour tree instances, can we compute/estimate some statistics among these contour tree instances? For example, what is the probability of two points p, q lying on an edge of the contour tree? What is the expected distance of two points p, q 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 [4]? We look into some of the computational challenges related to contour trees of terrains imposed by the uncertainty on the vertex heights.

Our contributions. The main results of this paper can be summarized as follows.

(A) We show (in Section 3) that the probability of two points p, q in $\mathbb{R}^2$ lying on an edge of the contour tree can be estimated in polylogarithmic time within additive error with high probability using a near-linear-size data structure. The results hold both for discrete and continuous distributions to represent the height of each terrain vertex.

(B) We define the distance of two points p, q on a contour tree 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 in [4]. We show (in Section 4) that two distance statistics, the expected distance of p, q in $\mathbb{R}^2$ and the probability that the distance of p, q is at least ℓ on the contour tree, can be estimated within additive error and/or relative error with high probability using a near-linear-size data structure.

(C) We show that (in Section 5) answering the above queries can be used for computing topological persistence and hydrology analysis in the presence of uncertainty.

(D) We present experimental results (in Section 6) to demonstrate the efficacy of our approach for estimating the probability of two points lying on an edge, and for estimating the distance statistics of two points.

Related work. The contour-tree problem under uncertainty has received some attention recently, see e.g. [10, 13, 14]. Kraus [13] 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 [14] 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. [10] 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 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.

Furthermore, there has been some work on terrain analysis in the presence of data uncertainty. The shortest-path problems have been studied on terrains in a probabilistic setting, termed as *uncertain terrains*, see e.g. [8, 9, 12]. Gray and Evans [9] showed that finding the *optimistic* shortest path on uncertain terrains is NP-hard, where for any vertex in the underlying triangulation defining an uncertain terrain, its height is represented as an interval and can take any value in the given interval, a path is characterized in $\mathbb{R}^2$, its length with respect to a terrain instance is defined in $\mathbb{R}^3$ by lifting it onto the terrain instance, and its length with respect to an uncertain terrain is defined to be the *minimum* length of this path among all terrains instances of the uncertain terrain. Later, Gary [8] extended the hardness result to the *pessimistic*-shortest-path problem where the length of a path with respect to an uncertain terrain is the *maximum* length of this path among all terrain instances of the uncertain terrain. Furthermore, Kholondyrev and Evans [12] showed that if we can walk only on the terrain edges, i.e., the (lifted) path is restricted on the terrain edges, then finding the pessimistic shortest path on uncertain terrains remains NP-hard and there exists a fully-polynomial time approximation scheme for it, while the optimistic version is polynomial-time solvable.

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 $\mathbb{M}$. 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.

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 connected component of $\mathbb{M}_\ell$ is called a *contour*. Each vertex v is contained in exactly one contour in $\mathbb{M}_{h(v)}$, which we call *the contour of v*. The contour of a local minimum or maximum v only consists of the single point v; the contour of a regular vertex is a simple polygonal cycle with non-empty interior; 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.

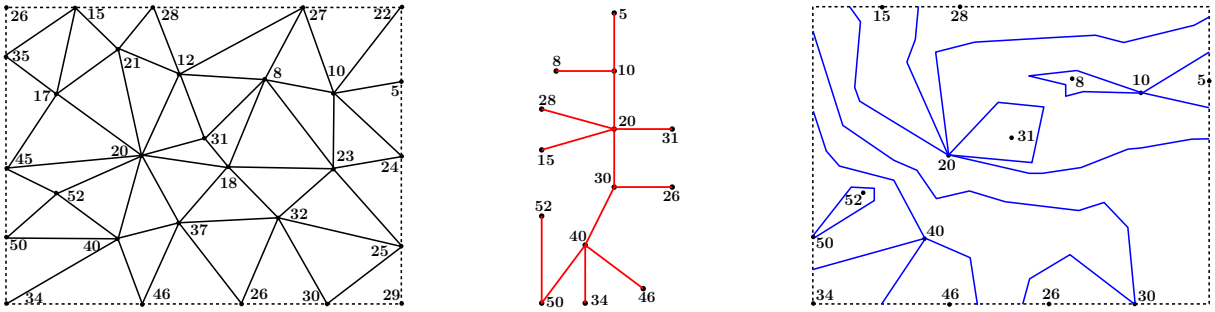
Formally, T_h is the quotient space in which each contour is represented by a point and connectivity is defined in terms of the quotient topology. Let $\rho : \mathbb{M} \to \mathsf{T}_h$ be the associated quotient map, which maps all points of a contour to a single point on an edge of T_h. Fix a point 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.[1] We will use h to denote the height function on the points of T_h as well.

The (closure of 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; if u or v is an extremal vertex, it is a simply connected region. These regions induce a planar subdivision, which we call the *height level map*[2] of Σ, and denote by M_h. See Fig. 1. 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.

Similarly, we define *extended height level map*. It is a subdivision induced by the level sets through all vertices of the triangulation, instead of the contours of critical vertices.

[1] This example figure is taken from [5], where each terrain has a bounding box, and a saddle vertex on the boundary is defined slightly different.

[2] The height level map was also defined in [5].



(a) (b) (c)

Figure 1. (a) A terrain Σ; (b) its contour tree T_h; (c) its height level map M_h.

The extended height level map has the same worst case complexity as that of 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 [4, 5].

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 saddle vertices of Σ.

Topological persistence. Topological persistence was introduced by Edelsbrunner et al. [7] 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 $\mathbb{M}_\ell$ while increasing ℓ. A component of $\mathbb{M}_\ell$ is started at a minimum vertex and ends at a saddle vertex when it joins with an older component. Similarly, a hole of $\mathbb{M}_\ell$ is started at a saddle vertex and ends at a maximum vertex. Based on this it is possible to define minimum-saddle and maximum-saddle persistence pairs between the critical vertex that starts a component or hole and the one that ends it. The persistence value of a persistence pair is simply the height difference between the vertices, i.e., it is the difference between the height at which the corresponding component was started and the height it was ended. Topological persistence is used to define the significance of various critical points.

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 uniform distribution and Gaussian distribution.

We will 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 $\boldsymbol{\Sigma}_\mathsf{H}, \mathsf{T}_\mathsf{H}$ 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 $\boldsymbol{\Sigma}_\mathsf{H}, \mathsf{T}_\mathsf{H}$ and M_H, respectively. We note that if H is a discrete distribution of description complexity k, then T_H and M_H can have $\Theta(k^n)$ size; see [19] 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 over the height functions, we wish to build a data structure that can quickly compute $\pi(p,q)$, 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 $|\pi(p,q) - \hat{\pi}(p,q)| \leq \varepsilon$ with probability at least $1-\delta$, for any $p, q \in \mathbb{R}^2$.

A Monte-Carlo algorithm. We fix a value $s \geq 1$, 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 $j \leq s$, using the algorithm in [5], 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 $\rho_j(p)$ and $\rho_j(q)$ lie on the same edge of the contour tree T_j.

Given two points $p, q \in \mathbb{R}^2$, for each $j \leq s$, 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. It remains to determine the value of s so that $|\hat{\pi}(p,q) - \pi(p,q)| \leq \varepsilon$ for all pairs of p and q in $\mathbb{R}^2$, with probability at least $1-\delta$.

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

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

Let Θ denote a set of representative pairs of points such that, if $|\hat{\pi}(p,q) - \pi(p,q)| \leq \varepsilon$ for all pairs $(p,q) \in \Theta$, then it 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 $|\hat{\pi}(p,q) - \pi(p,q)| \leq \varepsilon$ is at most $2|\Theta| \exp(-2\varepsilon^2 s)$. Hence, by setting $s = \frac{1}{2\varepsilon^2} \ln \frac{2|\Theta|}{\delta}$, $|\hat{\pi}(p,q) - \pi(p,q)| \leq \varepsilon$ 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 [15]. 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 over the height functions, and let M_H be the distribution over height level maps 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 $O(n^3 k^8)$.*

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 totally $O(nk^4)$ edges inside each triangle. The arrangement of these $O(nk^4)$ edges has complexity $O(n^2 k^8)$. Since we have n triangles, and edges are disjoint between any two triangles, therefore the complexity of the resulting planar map, which we denote by $\widetilde{\mathsf{M}}$, is $O(n^3 k^8)$. Furthermore, it is no hard to see that $\widetilde{\mathsf{M}}$ is finer than $\widehat{\mathsf{M}}$, hence $\widehat{\mathsf{M}}$ has complexity $O(n^3 k^8)$.

To construct $\widehat{\mathsf{M}}$, we need to remove redundant edges in $\widetilde{\mathsf{M}}$. First notice that each edge e in $\widetilde{\mathsf{M}}$ is determined by three vertices v_1, v_2, v_3 of a triangle in $\mathbb{M}$ and a saddle value h_v of some vertex v in $\mathbb{M}$. We test whether e is valid in $\widehat{\mathsf{M}}$, by checking the following two conditions:

(i) Whether it is possible for v to be a saddle vertex with saddle value h_v conditioning on the values that v_1, v_2, v_3 have taken. If v coincides with one of v_1, v_2, v_3, say v_1, then v_1 must take the same value as v. It can be checked in constant time assuming v is adjacent to constant number of triangles (hence vertices) in $\mathbb{M}$. If e is not valid, we remove e from $\widetilde{\mathsf{M}}$. We do this for every edge e in $\widetilde{\mathsf{M}}$.

(ii) Whether the edge e is able to reach v through a path of the same height as e. We can find all these edges in a breadth-first search manner starting from v, and then eliminate those edges e that cannot reach v through a path of the same height as e. Those eliminated edges are components of the level sets of $h(v)$ but do not lie in the same contour as v.

Finally, for every vertex v in $\mathbb{M}$, we check whether it has a chance to be a local minimum/maximum. If it does, we add v back into $\widetilde{\mathsf{M}}$ if v is removed in the phase of validating the edges in $\widetilde{\mathsf{M}}$. It is no hard to argue that the resulting map is $\widehat{\mathsf{M}}$. □

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

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 $|\hat{\pi}(p,q) - \pi(p,q)| \leq \varepsilon$ with probability at least $1 - \delta$.*

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 bounded number of distinct results. However, we just need to define a small set $\overline{\Theta}$ of pairs of points so that for any points $p, q \in \mathbb{R}^2$, there are two points $p', q' \in \overline{\Theta}$ such that $|\pi(p,q) - \pi(p',q')| \leq \varepsilon/2$. Then we can choose s large enough so that it permits at most $\varepsilon/2$ error on each pair of

points in $\overline{\Theta}$. Specifically, choosing $s = O((1/\varepsilon^2)\log(n|\overline{\Theta}|/\delta))$ is sufficient, so all that remains is to bound $|\overline{\Theta}|$.

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 $\overline{\mathsf{H}} = \{\overline{\mathsf{H}}_1, \ldots, \overline{\mathsf{H}}_n\}$ 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 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 of 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 of u and t, then $\rho_{\overline{h}}(p)$ (resp. $\rho_{\overline{h}}(q)$) lies on the edge of u (resp. t) and v, then again $\rho_{\overline{h}}(p)$, $\rho_{\overline{h}}(q)$ lie on the edge of u and t. Hence J_n has two connected components.

This concludes the proof. □

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 $j < n$. That is

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

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

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

Since $J_n(x_1, \ldots, x_{n-1})$ can be represented as at most two connected intervals, a classic result by Vapnik and Chervonenkis [18] (see also [11]) implies that

$$|\pi(p, q \mid x_1, \ldots, x_{n-1}) - \overline{\pi}(p, q \mid x_1, \ldots, x_{n-1})| \leq \alpha, \quad (2)$$

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

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

PROOF. Recall that

$$\pi(p, q) = \int \cdots \int \gamma_1(x_1) \cdots \gamma_n(x_n) \cdot \mathbf{1}(p, q; x_1, \ldots, x_n)\mathrm{d}x_n \ldots \mathrm{d}x_1.$$

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

$$\pi(p, q) \leq \frac{1}{\nu(\alpha)} \sum_{j_n=1}^{\nu(\alpha)} \int \cdots \int \gamma_1(x_1) \cdots \gamma_{n-1}(x_{n-1}) \cdot \mathbf{1}(p, q; x_1, \ldots, x_{n-1}, x_n^{j_n})\mathrm{d}x_{n-1} \ldots \mathrm{d}x_1 + \alpha.$$

Repeating this step $n - 1$ more times, we obtain

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

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

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

Finally, 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 two points $p, q \in \mathbb{R}^2$, in $O((1/\varepsilon^2)\log(n/(\varepsilon\delta))\log n)$ time, a value $\hat{\pi}(p, q)$ such that $|\hat{\pi}(p, q) - \pi(p, q)| \leq \varepsilon$ with probability at least $1 - \delta$.*

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. In particular, we are interested in the following two statistics: 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 $\Phi(p, q; \ell)$, i.e.,

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

As in Section 3, we use a simple Monte Carlo algorithm for estimating $\overline{\mathrm{d}}(p,q)$ and $\Phi(p,q;\ell)$. Namely, we fix a parameter $s \geq 1$. For each $i \leq s$, 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 $i \leq s$. 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 $\widehat{\Phi}(p,q;\ell) = |\{i \mid \mathrm{d}_{h_i}(p,q) \geq \ell\}|/s$ as an estimate of $\Phi(p,q;\ell)$.

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

Analysis for expected distance. For simplicity, we focus on the case when H is a discrete distribution. We begin by introducing a few definitions. For a vertex v_i, let $h_i^+ = \max_{1\leq j\leq k} h_i^j$, $h_i^- = \min_{1\leq j\leq k} h_i^j$, $\Delta_i = h_i^+ - h_i^-$; set $\Delta = \max_{1\leq i\leq n} \Delta_i$. 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 $i \leq n$, $h(v_i) \in [h_i^-, h_i^- + \Delta]$, therefore,

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

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

The following well-known lemma 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 $\varepsilon > 0$,*

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

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 $p, q \in \mathbb{R}^2$, Lemma 4.1 and Lemma 4.2 imply

$$\Pr[\mathsf{err}(p,q) > \varepsilon\Delta] \leq 2\exp\Big(-\frac{s\varepsilon^2}{2}\Big). \tag{3}$$

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 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 Δ_1, Δ_2 be two triangles in Ξ, and let $p_1, q_1 \in \Delta_1$, $p_2, q_2 \in \Delta_2$. 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)*, and *(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 with 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 Δ_x, Δ_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 Δ_x (resp. Δ_y), we can prove the following lemma. Its proof is rather tedious, so omitted from here.

LEMMA 4.4. *If $\mathsf{err}(p,q) \leq \varepsilon\Delta$ for every $(p,q) \in \Theta$, then $\mathsf{err}(p,q) \leq 3\varepsilon\Delta$ 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.*

Analysis for tail probability. Next we bound s, the number of samples, required for estimating $\Phi(p,q;\ell)$ 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

$$\Pr[|\Phi(p_0,q_0;\ell_0) - \widehat{\Phi}(p_0,q_0;\ell_0)| \geq \varepsilon] \leq 2\exp(-2\varepsilon^2/s). \tag{4}$$

As in Section 3, we construct a representative set Θ, where now $\Theta \subset \mathbb{R}^5$, so that if (4) holds for all triples in Θ, it also holds for all triples in $\mathbb{R}^5$.

We begin by constructing Ξ, the overlay of 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 $1 \leq i, j \leq k^3$, define a 4-variate linear function $g_{\zeta\eta}^{ij} : \mathbb{R}^4 \to \mathbb{R}$ as

$$g_{\zeta\eta}^{ij}(p,q) = h_\zeta^i(p) - h_\eta^j(q),$$

where $p, q \in \mathbb{R}^2$ and $h_\zeta^i \in \mathsf{H}_\zeta$, $h_\eta^j \in \mathsf{H}_\eta$. For each possible height h_l^r of vertex h_l, where $1 \leq l \leq n, 1 \leq r \leq k$, we define

$$g_{\zeta l}^{ir}(p,q) = h_\zeta^i(p) - h_l^r,$$

$$g_{\eta l}^{jr}(p,q) = h_\eta^j(q) - h_l^r.$$

Set

$$G_{\zeta\eta} = \{g^{ij}_{\zeta\eta}, -g^{ij}_{\zeta\eta}, g^{ir}_{\zeta l}, g^{jr}_{\eta l} \mid 1 \le i, j \le k^3, 1 \le l \le n, 1 \le r \le k\}.$$

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 book [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 $\Phi(p, q; \ell) = \Phi(p_\tau, q_\tau; \ell_\tau)$.*

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. For simplicity, assume both extremal points of $\mathrm{d}_h(p_0, q_0)$ are its endpoints and $h(p_0) \ge h(q_0)$; the argument of 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 $h^j_\eta \in \mathsf{H}_\eta$. Then $\mathrm{d}_h(p_0, q_0) = h^i_\zeta(p_0) - h^j_\eta(q_0) = g^{ij}_{\zeta\eta}(p_0, q_0) \ge \ell_0$. 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. □

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 $\Phi(p, q; \ell)$ within additive error ε with probability at least $1 - \delta$.*

5. HYDROLOGY ANALYSIS QUERIES

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.

6.1 Experimental setup

Datasets. We use the regular terrain dataset San Bernardino, which is originally a grid composed of $128 \times 128 = 16384$ vertices. We use the upper-left quarter of the data[3], 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. 1(a), and the other based on a grid of size $17 \times 17 = 289$ vertices with randomly-generated (exact) heights. The results are pretty similar to each other, and we choose to present the San Bernardino dataset by default. All these datasets do not have uncertainty on the vertex heights, and we introduce uncertainty below.

Uncertainty. Given a parameter Δ, we introduce uncertainty on the vertex heights as follows. For any vertex v, and let h_v denote its height in the above dataset (without uncertainty). We consider both continuous and discrete distributions describing the vertex heights. For continuous distributions, we consider uniform continuous distribution $U(h_v - \Delta/2, h_v + \Delta/2)$, Gaussian distribution $N(h_v, \Delta/6)$, and bimodal Gaussian distribution characterized by $N(h_v - \Delta/4, \Delta/12)$ and $N(h_v + \Delta/4, \Delta/12)$. For discrete distributions, we assume that each vertex has k possible heights drawn randomly from the uniform distribution $U(h_v - \Delta/2, h_v + \Delta/2)$, each with some probability being chosen. We consider weighted discrete distribution where all the k heights have different (randomly-generated) probabilities being chosen, and uniform discrete distribution where all the k heights have the same probability $1/k$ being chosen. In the experiments, we set $k = 5$ and $\Delta \in \{10, 20, 30, 40, 50\}$.

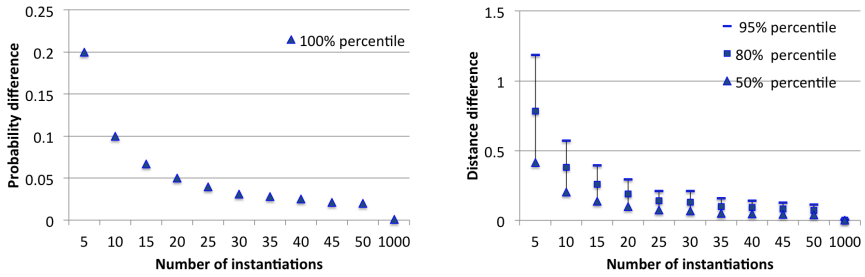
Queries. Among all $t = 7938$ triangles in the underlying triangulation of the dataset, we randomly select 100 of them. For each selected triangle, we choose its centroid as the representative point. Each two distinct representative points p, q constitute one query (p, q), resulting in $\binom{100}{2}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

Obtaining the exact values of $\pi(p, q)$ and $\mathsf{Ed}(p, q)$. Unfortunately, we do not know how to compute $\pi(p, q)$ and $\mathsf{Ed}(p, q)$ exactly. 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).

Measuring the convergence. For each query (p, q) and the j-th round, we compute the probability 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 100% percentile (i.e., the maximum) of these probability differences. Similarly, we can look at the distance differences, and we report the 50%, 80% and 95% percentiles of these distance differences.

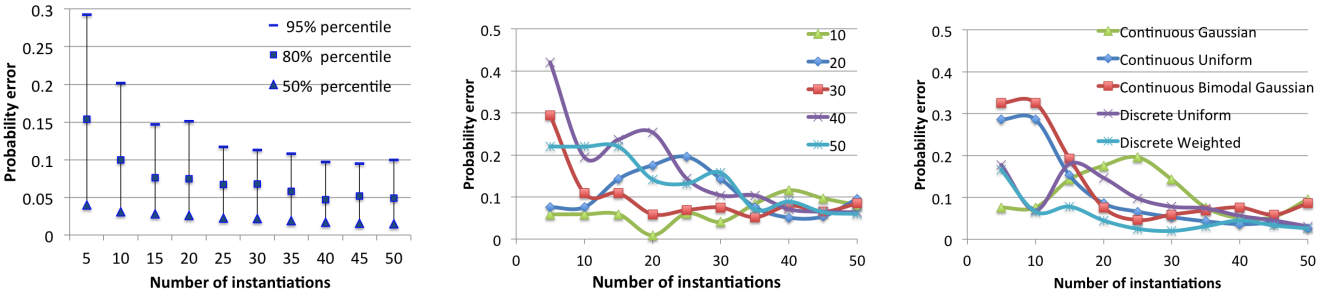
Measuring the effectiveness. For each query (p, q), we compute the probability error $|\pi(p, q) - \hat{\pi}(p, q)|$, and the distance error $|\mathsf{Ed}(p, q) - \widehat{\mathsf{Ed}}(p, q)|$. Among all $\binom{100}{2} = 4950$ queries, we report the 50%, 80% and 95% percentiles of these errors. Note that if $\pi(p, q) = 0$, then $\hat{\pi}(p, q) = 0$ and the error is 0. It also holds for the case when $\pi(p, q) = 1$. We

[3]We do this simply for avoiding memory overconsumption.



(a) (b)

Figure 2. Probability difference of $\pi(p, q)$, and distance difference of $\mathsf{Ed}(p, q)$.



(a) (b) (c)

Figure 3. Probability errors of $\pi(p, q)$.

rule out such queries when we measure the probability errors (otherwise the percentile errors are even smaller).

6.2 Experimental results

Convergence of our methods. We tested how our methods converged as we varied s, the number of instantiations. Fig. 2(a)–(b) show the probability differences and distance differences decrease very fast, and they become very close to 0 when $s = 1000$. Therefore, we take the estimates using 1000 samples as exact values. For Fig. 2(a)–(b), we used Gaussian distribution and $\Delta = 20$.

Probability errors in estimating $\pi(p, q)$. We tested how the probability errors changed as we varied s, the number of instantiations. Fig. 3(a)–(c) show the probability errors as s varies from 5 to 50. Not surprisingly, as s increases, the probability errors decrease, and the errors are reasonably small when $s \geq 20$. The smaller uncertainty (as denoted by Δ) we have, the smaller error we tend to have, though Fig. 3(b) does not convey a very clear pattern. Regarding the distributions, the discrete weighted distribution outperformed others, but it may be just the case we randomly generated "good" weights for this particular run. For Fig. 3(a), we used Gaussian distribution and $\Delta = 20$. For Fig. 3(b), we used Gaussian distribution, $s = 20$, and 95% percentile. For Fig. 3(c), we used $\Delta = 20$ and 95% percentile.

Distance errors in estimating $\mathsf{Ed}(p, q)$. Similarly, Fig. 4(a)–(c) show that, as the number of instantiations s increases, the distance errors decrease, and the distance error is far smaller than the uncertainty level Δ. The smaller uncertainty we have, the smaller error we have, as Fig. 4(b) indicates clearly. Regarding the distributions, Gaussian distribution and discrete distributions outperformed other distributions. Fig. 4(a)–(c) have the same settings as Fig. 3(a)–(c).

Errors vs exact values. We tested how errors are related to their exact values. Fig. 5(a)-(b) show that both the probability errors and the distance errors are independent with their underlying exact values. For Fig. 5, we used the smallest dataset as shown in Fig. 1(a), Gaussian distribution, $s = 20$, and $\Delta = 10$. The number of queries is $\binom{40}{2} = 780$, as the smallest dataset has only 40 triangles. Note that Fig. 5(a) looks sparser than Fig. 5(b), since there are many (1, 0) and (0, 0) points in Fig. 5(a).

Distribution of probability values. We tested how the percentage of queries (p, q) with $\pi(p, q) = 0$, $\pi(p, q) = 1$ and $\pi(p, q) \in (0, 1)$ varied as we increased the uncertainty level. 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 query with $\pi(p, q) = 0$ decreases very slightly. Fig. 5(c) has the same setting as Fig. 5(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 an 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 and/or relative error. We also conducted some preliminary experiments to demonstrate the effectiveness of our methods. We conclude this paper with some open problems:

1. 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?

2. What is a robust and useful contour tree representation of a terrain in the presence of data uncertainty?

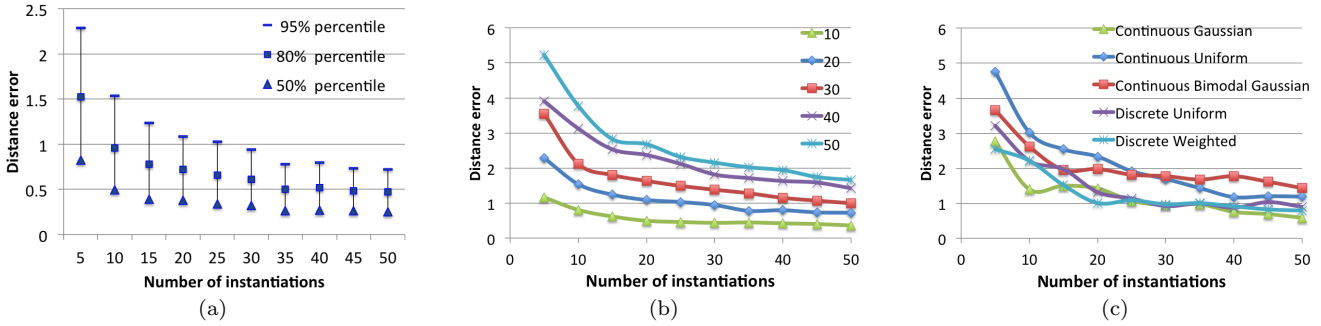


Figure 4. Distance errors of $\mathsf{Ed}(p, q)$.

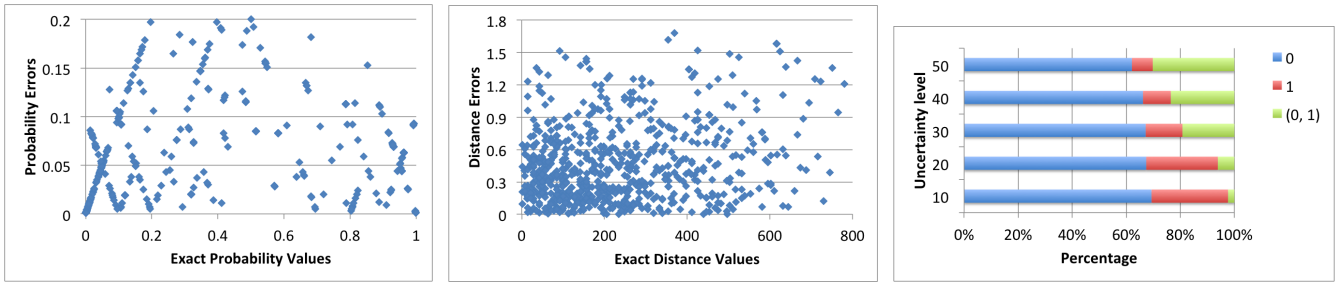


Figure 5. (a)-(b) Errors vs exact values. (c) The percentage of queries (p, q) with $\pi(p, q) = 0$, $\pi(p, q) = 1$ and $\pi(p, q) \in (0, 1)$ vs uncertainty level.

Acknowledgements. P. Agarwal and W. Zhang are supported by NSF under grants CCF-09-40671, CCF-10-12254, and CCF-11-61359, by ARO grants W911NF-07-1-0376 and W911NF-08-1-0452, and by an ERDC contract W9132V-11-C-0003. The authors would also like to acknowledge 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. *CoRR*, abs/1406.4005, 2014.

[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(1):11:1–11:21, 2010.

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

[4] U. Bauer, X. Ge, and Y. Wang. Measuring distance between reeb graphs. In *Proc. 30th Annu. ACM Sympos. Comput. Geom.*, pages 464–473, 2014.

[5] M. d. Berg and M. J. v. Kreveld. Trekking in the alps without freezing or getting tired. In *Proc. 1st Annu. European Sympos. Algs.*, pages 121–132, 1993.

[6] H. Carr, J. Snoeyink, and U. Axen. Computing contour trees in all dimensions. In *Proc. 11th Annu. ACM-SIAM Sympos. Discrete Algs.*, pages 918–926, 2000.

[7] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *Proc. 41th Annu. IEEE Sympos. Found. Comput. Sci.*, pages 454–463, 2000.

[8] C. Gray. Shortest paths on uncertain terrains. Master's thesis, Department of Computer Science, University of British Columbia, 2004.

[9] C. Gray and W. S. Evans. Optimistic shortest paths on uncertain terrains. In *Proc. 16th Canad. Conf. Comput. Geom.*, pages 68–71, 2004.

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

[11] S. Har-Peled. *Geometric Approximation Algorithms*. Mathematical Surveys and Monographs. American Mathematical Society, 2011.

[12] Y. Kholondyrev and W. Evans. Optimistic and pessimistic shortest paths on uncertain terrains. In *Proc. 19th Canad. Conf. Comput. Geom.*, pages 197–200, 2007.

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

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

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

[16] S. P. Tarasov and M. N. Vyalyi. Construction of contour trees in 3D in $O(n \log n)$ steps. In *Proc. 14th Annu. ACM Sympos. Comput. Geom.*, pages 68–75, 1998.

[17] 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 Annu. ACM Sympos. Comput. Geom.*, pages 212–220, 1997.

[18] 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.

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