The Hunting of the Bump: On Maximizing Statistical Discrepancy

Deepak Agarwal* Jeff M. Phillips† Suresh Venkatasubramanian‡

Abstract

Anomaly detection has important applications in biosurveilance and environmental monitoring. When comparing measured data to data drawn from a baseline distribution, merely, finding clusters in the measured data may not actually represent true anomalies. These clusters may likely be the clusters of the baseline distribution. Hence, a discrepancy function is often used to examine how different measured data is to baseline data within a region. An anomalous region is thus defined to be one with high discrepancy.

In this paper, we present algorithms for maximizing statistical discrepancy functions over the space of axis-parallel rectangles. We give provable approximation guarantees, both additive and relative, and our methods apply to any convex discrepancy function. Our algorithms work by connecting statistical discrepancy to combinatorial discrepancy; roughly speaking, we show that in order to maximize a convex discrepancy function over a class of shapes, one needs only maximize a linear discrepancy function over the same set of shapes.

We derive general discrepancy functions for data generated from a one- parameter exponential family. This generalizes the widely-used Kulldorff scan statistic for data from a Poisson distribution. We present an algorithm running in $O(\frac{1}{\epsilon}n^2 \log^2 n)$ that computes the maximum discrepancy rectangle to within additive error ϵ, for the Kulldorff scan statistic. Similar results hold for relative error and for discrepancy functions for data coming from Gaussian, Bernoulli, and gamma distributions. Prior to our work, the best known algorithms were exact and ran in time $O(n^4)$.

1 Introduction

Outlier detection (or "bump hunting"[6]) is a common problem in data mining. Unlike in robust clustering settings, where the goal is to detect outliers in order to remove them, outliers are viewed as *anomalous events* to be studied further. In the area of biosurveillance for example, an outlier would consist of an area that had an unusually high disease rate (disease occurence per unit population) of a particular ailment. In environmental monitoring scenarios, one might monitor the rainfall over an area and wish to determine whether any region had unusually high rainfall in a year, or over the past few years.

A formal statistical treatment of these problems allows us to abstract them into a common framework. Assume that data (disease rates, rainfall measurements, temperature) is generated by some stochastic spatial process. Points in space are either fixed or assumed to be generated from some spatial point process and measurements on points are assumed to be statistically independent and follow a distribution from a one-parameter exponential family. Also, let $b(\cdot)$ be some baseline measure defined on the plane. For instance, $b(\cdot)$ can be the counting measure (counts the number of points in a region), volume measure (measures volume of a region), weighted counting measure (adds up non-negative weights attached to points in a region). In biosurveillance, the counting measure (gives the region population) is often used to discover areas with elevated disease risk. Weighted counting measures which aggregate weights assigned to points based on some attribute (e.g. race of an individual) have also been used (see [13] for an example). Let p be a set of points generating a set of measurements $m(p)$. Given a measure of discrepancy f that takes as input the functions $m(\cdot)$ and $b(\cdot)$ and a collection of regions $\mathcal{S}$, the problem of *statistical discrepancy* can be defined as:

Find the region $S \in \mathcal{S}$ for which f invoked on the measurements for points in S is maximized.

Statistical discrepancy functions arise by asking the following question: "How likely is it that the observed points in S come from a distribution that is different than the distribution of points in S^c?". The function f is derived using a *likelihood ratio test* which has high statistical power to detect the actual location of clusters, but is computationally difficult to deal with. As a consequence, most algorithmic work on this problem has focused either on fast heuristics that do not search the entire space of shapes, or on conservative heuristics that guarantee finding the maximum discrepancy region and will often (though not always) run in time less than the worst-case bound of $|\mathcal{S}|$ times the function evaluation cost.

Apart from identifying the region for which f is maximized, it is customary to evaluate the likelihood of the identified cluster being generated by chance, i.e., compute the probability (called p-value) of maximum

*AT&T Labs – Research dagarwal@research.att.com

†Duke University jeffp@cs.duke.edu, supported by a James B. Duke Graduate Fellowship and a NSF Graduate Research Fellowship.

‡AT&T Labs – Research suresh@research.att.com

SODA '06, January 22-26, Miami, FL
 ISBN 0-89871-605-5/06/01

discrepancy to exceed the observed maximum discrepancy under the null hypothesis of no clustering effect. A small p-value (e.g. $< .05$) will mean the identified cluster is statistically significant. Since the distribution of f is usually not analytically tractable, randomization tests ([12, 7]) which involve multiple instances of the maximum discrepancy computation are run for data generated from the null model. Thus, the computation of statistical discrepancy is the main algorithmic bottleneck and is the problem we focus on in this paper.

1.1 Our Contributions In this paper, we present algorithms with non-trivial worst-case running time bounds for approximating a variety of statistical discrepancy functions. Our main result is a structural theorem that reduces the problem of maximizing any convex discrepancy function over a class of shapes to maximizing a simple linear discrepancy function over the same class of shapes.

The power of this theorem comes from the fact that there are known algorithms for maximizing special kinds of linear discrepancy functions, when the class of shapes consists of axis-parallel rectangles. Given two sets of red and blue points in the plane, the *combinatorial discrepancy* of a region is the absolute difference between the number of red and blue points in it. Combinatorial discrepancy is very valuable when derandomizing geometric algorithms; it also appears in machine learning as the relevant function for the *minimum disagreement problem*, where red and blue points are thought of as good and bad examples for a classifier, and the regions are hypotheses. This problem (and a more general variant of it) was considered by Dobkin, Maass and Gunopoulos in 1995 [5], where they showed that combinatorial discrepancy for axis-parallel rectangles in the plane could be maximized exactly in time $O(n^2 \log n)$, far better than the $O(n^4)$ bound that a brute-force search would entail.

We show that the Dobkin *et. al.* algorithm can be extended fairly easily to work with general linear discrepancy functions. This result, combined with our general theorem, allows us to approximate *any* convex discrepancy function over the class of axis-parallel rectangles. We summarize our results in Table 1; as an example, we present an additive approximation algorithm for the Kulldorff scan statistic that runs in time $O(\frac{1}{\epsilon} n^2 \log^2 n)$, which compares favorably to the (trivial) $O(n^4)$ running time to compute an exact solution.

Essentially, the reduction we use allows us to decouple the measure of discrepancy (which can be complex) from the shape class it is maximized over. Using our approach, if you wanted to maximize a general discrepancy function over a general shape class, you need only consider combinatorial discrepancy over this class. As a demonstration of the generality of our method, we also present algorithms for approximately maximizing discrepancy measures that derive from different underlying distributions. In fact, we provide general expressions for the one-parameter exponential family of distributions which includes Poisson, Bernoulli, Gaussian and Gamma distributions. For the Gaussian distribution, the measure of discrepancy we use is novel, to the best of our knowledge. It is derived from maximum likelihood considerations, has a natural interpretation as a χ^2 distance, and may be of independent interest.

Another notion of outlier detection incorporates a time dimension. In *prospective outlier detection*, we would like to detect the maximum discrepancy region over all time intervals starting from the present and going backwards in time. We show that linear discrepancy can be maximized over such time intervals and, as a consequence of our reduction, show that any convex discrepancy function can be approximately maximized.

2 Related Work

Detecting clustering effects in spatial data is a well-studied problem in statistics[1]. Much of the early focus has been on devising efficient statistical tests to detect presence of clustering at a global level without emphasis on identifying the actual clusters (see [3, Chapter 8]). The spatial scan statistic, introduced by Kulldorff [13] provides an elegant solution for detection and evaluation of spatial clusters. The technique has found wide applicability in areas like public health, biosurveillance, environmental monitoring *etc.* For interesting applications and detailed description of scan statistics, we refer the reader to [9, 10]. Generalization of the spatial scan statistic to a space-time scan statistic for the purpose of prospective surveillance has been proposed by Kulldorff [14], and Iyengar [8] suggested the use of "expanding-in-time" regions to detect space-time clusters. We note that the algorithms described by Kulldorff are heuristics: they do not guarantee any bound on the quality of the solution, and do not traverse the entire space of regions. The regions he considers are circular, and cylindrical in the case of prospective surveillance.

Dobkin and Eppstein [4] were the first to study efficient algorithms to compute maximum discrepancy over a range space. Their algorithms compute discrepancy in

[1] It goes without saying that there is a huge literature on clustering spatial data. Since our focus is primarily on statistically sound measures, a survey of general clustering methods is beyond the scope of this paper.

	This paper		Prior work
	OPT $-\epsilon$	OPT$/(1+\epsilon)$	Exact
Poisson (Kulldorff)/Bernoulli/Gamma	$O(\frac{1}{\epsilon}n^2 \log^2 n)$	$O(\frac{1}{\epsilon}n^2 \log^2 n)$	$O(n^4)$
Gaussian	$O(\frac{1}{\epsilon}n^3 \log n \log\log n)$	$O(\frac{1}{\epsilon}n^2 \log^2 n)$	$O(n^4)$

Table 1: Our results. For prospective discrepancy, multiply all bounds by n, and for higher dimensions, multiply by n^{2d-4}.

a region R as a difference between the fraction of points in R and the fraction of the total area represented by R. This measure stems from evaluating fundamental operations for computer graphics. Their ranges were half spaces and axis-oriented orthants centered at the origin, limited to the unit cube, and their results extended to d-dimensional spaces. Subsequently Dobkin, Gunopulous, and Maass [5] developed algorithms for computing maximum bichromatic discrepancy over axis-parallel rectangular regions, where the bichromatic discrepancy of a region is the difference between the number of red points and the number of blue points in the region. This solves the *minimum disagreement problem* from machine learning, where an algorithm finds the region with the most *good* points and the fewest *bad* points, a key subroutine in agnostic PAC learning.

Recently, Neill and Moore have developed a series of algorithms to maximize discrepancy for measures such as the Kulldorff spatial scan statistic. Their approach works for axis parallel squares [15] and rectangles [16]. Their algorithms are conservative, in that they always find the region of maximum discrepancy. The worst-case running time of their algorithms is $O(n^4)$ for rectangles and $O(n^2)$ for fixed-size squares since the algorithms enumerate over all valid regions. However, they use efficient pruning heuristics that allow for significant speedup over the worst case on most data sets. An alternative approach by Friedman and Fisher [6] greedily computes a high discrepancy rectangle, but has no guarantees as to how it compares to the optimal. Their approach is quite general, and works in arbitrary dimensional spaces, but is not conservative: many regions will remain unexplored.

A one-dimensional version of this problem has been studied in bioinformatics [11]. The range space is now the set of all intervals, a problem with much simpler geometry. In this setting, a relative ϵ-approximation can be found in $O(\frac{1}{\epsilon^2}n)$ time.

A related problem that has a similar flavor is the so-called *Heavy Hitters* problem [2, 1]. In this problem, one is given a multiset of elements from a universe, and the goal is to find elements whose frequencies in the multiset are unusually high (i.e much more than the average). In a certain sense, the heavy hitter problem fits in our framework if we think of the baseline data as the uniform distribution, and the counts as the measurements. However, there is no notion of ranges[2] and the heavy hitter problem itself is interesting in a streaming setting, where memory is limited; if linear memory is permitted, the problem is trivial to solve, in contrast to the problems we consider.

3 Preliminaries

Let P be a set of n points in the plane. Measurements and baseline measures over P will be represented by two functions, $m : P \to \mathbb{R}$ and $b : P \to \mathbb{R}$. $\mathcal{R}$ denotes a range space over P. A *discrepancy function* is defined as $d : (m, b, R) \to \mathbb{R}$, for $R \in \mathcal{R}$.

Let $m_R = \sum_{p\in R} m(p)/M, b_R = \sum_{p\in R} b(p)/B$, where $M = \sum_{p\in U} m(p)$, $B = \sum_{p\in U} b(p)$, and U is some box enclosing all of P. *We will assume that d can be written as a convex function of m_R, b_R.* All the discrepancy functions that we consider in this paper satisfy this condition; most discrepancy functions considered prior to this work are convex as well. We can write $d(m, b, R)$ as a function $d' : [0, 1]^2 \to \mathbb{R}$, where $d(m, b, R) = d'(m_R, b_R)$. We will use d to refer to either function where the context is clear.

Linear discrepancy functions are a special class of discrepancy functions where $d = \alpha \cdot m_R + \beta \cdot b_R + \gamma$. It is easy to see that combinatorial (bichromatic) discrepancy, the difference between the number of red and blue points in a region, is a special case of a linear discrepancy function.

The main problem we study in this paper is:

PROBLEM 3.1. (MAXIMIZING DISCREPANCY) *Given a point set P with measurements m, baseline measure b, a range space $\mathcal{R}$, and a convex discrepancy function d, find the range $R \in \mathcal{R}$ that maximizes d.*

An equivalent formulation, replacing the range R by the point $r = (m_R, b_R)$ is:

PROBLEM 3.2. *Maximize convex discrepancy function d over all points $r = (m_R, b_R), R \in \mathcal{R}$.*

[2]Hierarchical heavy hitters provide the beginnings of such a notion.

Assume that points now arrive with a timestamp $t(\cdot)$, along with the measurement $m(\cdot)$ and baseline $b(\cdot)$. In *prospective discrepancy* problems, the goal is to maximize discrepancy over a region in space and time defined as $R \times [t, t_{\text{now}}]$, where R is a spatial range. In other words, the region includes all points with a timestamp between the present time and some time t in the past. Such regions are interesting when attempting to detect *recent* anomalous events.

PROBLEM 3.3. (PROSPECTIVE DISCREPANCY) *Given a point set P with measurements m, baseline measure b, timestamps t, a range space $\mathcal{R}$, and a convex discrepancy function d, find the range $T = (R, [t^*, \infty]), R \in \mathcal{R}$ that maximizes d.*

3.1 Boundary Conditions As we shall see in later sections, the discrepancy functions we consider are expressed as log-likelihood ratios. As a consequence, they tend to ∞ when either argument tends to zero (while the other remains fixed). Another way of looking at this issue is that regions with very low support often correspond to overfitting and thus are not interesting. Therefore, this problem is typically addressed by requiring a *minimum level of support* in each argument. Specifically, we will only consider maximizing discrepancy over shapes R such that $m_R > C/n, b_R > C/n$, for some constant C. In mapping shapes to points as described above, this means that the maximization is restricted to points in the square $S_n = [C/n, 1 - C/n]^2$. For technical reasons, we will also assume that for all p, $m(p), b(p) = \Theta(1)$. Intuitively this reflects the fact that measurement values are independent of the number of observations made.

4 A Convex Approximation Theorem

We start with a general approximation theorem for maximizing a convex discrepancy function d. Let $\ell(x, y) = a_1 x + a_2 y + a_3$ denote a linear function in x and y. Define an *ϵ-approximate family* of d to be a collection of linear functions $\ell_1, \ell_2, \ldots, \ell_t$ such that $l^U(x, y) = \max_{i \le t} \ell_i(x, y)$, the *upper envelope* of the ℓ_i, has the property that $l^U(x, y) \le d(x, y) \le l^U(x, y) + \epsilon$ Define a *relative* ϵ-approximate family of d to be a collection of linear functions $\ell_1, \ell_2, \ldots, \ell_t$ such that $l^U(x, y) \le d(x, y) \le (1 + \epsilon) l^U(x, y)$

LEMMA 4.1. *Let $\ell_1, \ell_2, \ldots, \ell_t$ be an ϵ-approximate family of a convex discrepancy function $d : [0, 1]^2 \to \mathbb{R}$. Consider any point set $\mathcal{C} \subset [0, 1]^2$. Let $(x_i^*, y_i^*) = \arg\max_{\mathbf{p} \in \mathcal{C}} \ell_i(\mathbf{p}_x, \mathbf{p}_y)$, and let $(x^*, y^*) = \arg\max_{x_i^*, y_i^*} \ell_i(x_i^*, y_i^*)$. Let $d^* = \sup_{\mathbf{p} \in \mathcal{C}} d(\mathbf{p}_x, \mathbf{p}_y)$, $d_{\text{inf}} = \inf_{\mathbf{q} \in [0,1]^2} d(\mathbf{q}_x, \mathbf{q}_y)$ and let $m = \max(l^U(x^*, y^*), d_{\text{inf}})$. Then*

$$m \le d^* \le m + \epsilon$$

If $\ell_1, \ell_2, \ldots, \ell_t$ is a relative ϵ-approximate family, then

$$m \le d^* \le (1 + \epsilon) m$$

Proof Sketch. By construction, each point $(x_i^*, y_i^*, l_i(x_i^*, y_i^*))$ lies on the upper envelope l^U. The upper envelope is convex, and lower bounds $d(\cdot)$, and therefore in each *patch* of l^U corresponding to a particular ℓ_i, the maximizing point (x_i^*, y_i^*) also maximizes $d(x, y)$ in the corresponding patch. This is only false for the patch of l^U that supports the minimum of $d(x, y)$, where the term involving d_{inf} is needed. This corresponds to adding a single extra plane tangent to $d(\cdot)$ at its minimumm, which is unique by virtue of $d(\cdot)$ being convex.

LEMMA 4.2. *Let $f : [0, 1]^2 \to \mathbb{R}$ be a convex smooth function. Let $\tilde{f} : [0, 1]^2 \to \mathbb{R}$ be the linear approximation to f represented by the hyperplane tangent to f at $\mathbf{p} \in [0, 1]^2$. Then $\tilde{f}(\mathbf{p}) \le f(\mathbf{p})$, and $f(\mathbf{p}) - \tilde{f}(\mathbf{q}) \le \|\mathbf{p} - \mathbf{q}\|^2 \lambda^*$, where λ^* is the maximum value of the largest eigenvalue of $H(f)$, maximized along the line joining $\mathbf{p}$ and $\mathbf{q}$.*

Proof. $\tilde{f}(\mathbf{q}) = f(\mathbf{p}) + (\mathbf{q} - \mathbf{p})^\top \nabla f(\mathbf{p})$. The inequality $\tilde{f}(\mathbf{p}) \le f(\mathbf{p})$ follows from the convexity of f. By Taylor's theorem for multivariate functions, the error $f(\mathbf{p}) - \tilde{f}(\mathbf{q}) = (\mathbf{q} - \mathbf{p})^\top H(f)(\mathbf{p}^*)(\mathbf{q} - \mathbf{p})$, where $H(f)$ is the Hessian of f, and $\mathbf{p}^*$ is some point on the line joining $\mathbf{p}$ and $\mathbf{q}$.

Writing $\mathbf{q} - \mathbf{p}$ as $\|\mathbf{q} - \mathbf{p}\| \hat{\mathbf{x}}$, where $\hat{\mathbf{x}}$ is a unit vector, we see that the error is maximized when the expression $\hat{\mathbf{x}}^\top H(f) \hat{\mathbf{x}}$ is maximized, which happens when $\hat{\mathbf{x}}$ is the eigenvector corresponding to the largest eigenvalue λ^* of $H(f)$.

Let $\lambda^* = \sup_{\mathbf{p} \in S_n} \lambda_{\max}(H(f)(\mathbf{p}))$. Let $\epsilon_{\mathbf{p}}(\mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|^2 \lambda^*, \epsilon_{\mathbf{p}}^R(\mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|^2 \lambda^* f(\mathbf{p})$.

LEMMA 4.3. *Let $\mathcal{C} \subset S_n$ be a set of t points such that for all $\mathbf{q} \in S_n, \min_{\mathbf{p} \in \mathcal{C}} \epsilon_{\mathbf{p}}(\mathbf{q})$(resp. $\epsilon_{\mathbf{p}}^R(\mathbf{q})$) $\le \epsilon$. Then the t tangent planes at the points $f(\mathbf{p}), \mathbf{p} \in \mathcal{C}$ form an ϵ-approximate (resp. relative ϵ-approximate) family for f.*

Proof. Let $\mathcal{C} = \{\mathbf{p}_1, \ldots, \mathbf{p}_t\}$. Let l_i denote the tangent plane at $f(\mathbf{p}_i)$. For all i, $l_i(\mathbf{q}) \le f(\mathbf{q})$ by Lemma 4.2. Let $j = \arg\min_i \epsilon_{\mathbf{p_i}}(\mathbf{q})$. Then $l_j(\mathbf{q}) = \max_i l_i(\mathbf{q}) \le \epsilon$. A similar argument goes through for $\epsilon_{\mathbf{p_i}}^R(\mathbf{q})$

The above lemmas indicate that in order to construct an ϵ-approximate family for the function f, we

need to sample an appropriate set of points from S_n. A simple area-packing bound, using the result from Lemma 4.3, indicates that we would need $O(\lambda^*/\epsilon)$ points (and thus that many planes). However, λ^* is a function of n. A stratified grid decomposition can exploit this dependence to obtain an improved bound.

THEOREM 4.1. *Let $f : [0,1]^2 \to \mathbb{R}$ be a convex smooth function, and fix $\epsilon > 0$. Let $\lambda(n) = \lambda^*(S_n)$. Let $F(n,\epsilon)$ be the size of an ϵ-approximate family for f, and let $F^R(n,\epsilon)$ denote the size of a relative ϵ-approximate family. Let $\lambda(n) = O(n^d)$. Then,*

$$F(n,\epsilon) = \begin{cases} O(1/\epsilon) & d = 0 \\ O(\frac{1}{\epsilon}\log_{\frac{1}{d}}\log n) & 0 < d < 1 \\ O(\frac{1}{\epsilon}\log n) & d = 1 \\ O(\frac{1}{\epsilon}n^{d-1}\log_d \log n) & d > 1 \end{cases}$$

Let $\lambda'(n) = \lambda(n)/f_{\max}(n)$, where $f_{\max}(n)$ denotes $\max_{\mathbf{p}\in S_n} f(\mathbf{p})$. Then $F^R(n,\epsilon)$ has size chosen from the above cases according to $\lambda'(n)$.

Proof. The relation between $F^R(n,\epsilon)$ and $F(n,\epsilon)$ follows trivially from the relationship between $\epsilon^R_{\mathbf{p}}(\mathbf{q})$ and $\epsilon_{\mathbf{p}}(\mathbf{q})$.

If $\lambda(n)$ is $O(1)$, then λ^* can be upper bounded by a constant, resulting in an ϵ-approximate family of size $O(1/\epsilon)$. The more challenging case is when λ^* is an increasing function of n.

Suppose $\lambda^* = O(n^d)$ in the region S_n. Consider the following adaptive gridding strategy. Fix numbers $n_0, n_1, \ldots n_k$, $n_{k+1} = n$. Let $A_0 = S_{n_0} = [1/n_0, 1 - 1/n_0]^2$. Let $A_i = S_{n_i} - \cup_{j<i} A_i$. Thus, A_0 is a square of side $1 - 2/n_0$, and each A_i is an annulus lying between $S_{n_{i+1}}$ and S_{n_i}. A_0 has area $O(1)$ and each $A_i, i > 0$ has area $O(1/n_{i-1})$. In each region A_i, $\lambda^*(A_i) = O(n_i^d)$.

How many points do we need to allocate to A_0? A simple area bound based on Lemma 4.3 shows that we need $\lambda^*(A_0)/\epsilon$ points, which is $O(n_0^d/\epsilon)$. In each region A_i, a similar area bound yields a value of $O(n_i^d/\epsilon n_{i-1})$. Thus the total number of points needed to construct the ϵ-approximate family is $N(d,k) = n_0^d/\epsilon + \sum_{0<i\le k+1} n_i^d/\epsilon n_{i-1}$.

Balancing this expression by setting all terms equal, and setting $l_i = \log n_i$, we obtain the recurrence

$$l_i = \frac{(d+1)l_{i-1} - l_{i-2}}{d} \tag{4.1}$$

$$l_1 = \frac{d+1}{d} l_0 \tag{4.2}$$

CLAIM 4.1. $l_{k+1} = \log n = (1 + \sum_{i=1}^{j} d^{-i}) l_{k-j+1} - (\sum_{i=1}^{j} d^{-i}) l_{k-j}$

Proof. The proof is by induction. The statement is true for $j = 1$ from Eq.(4.1). Assume it is true upto j. Then

$$\begin{aligned} l_{k+1} &= \left(1 + \sum_{i=1}^{j} d^{-i}\right) l_{k-j+1} - \left(\sum_{i=1}^{j} d^{-i}\right) l_{k-j} \\ &= \left(1 + \sum_{i=1}^{j} d^{-i}\right) \left[\frac{(d+1)l_{k-j} - l_{k-j-1}}{d}\right] - \\ &\quad \left(\sum_{i=1}^{j} d^{-i}\right) l_{k-j} \\ &= \left(1 + \sum_{i=1}^{j+1} d^{-i}\right) l_{k-j} - \left(\sum_{i=1}^{j+1} d^{-i}\right) l_{k-j-1} \end{aligned}$$

Setting $j = k$ in Claim 4.1 yields the expression $\log n = (1 + \sum_{i=1}^{k} d^{-i}) l_1 - (\sum_{i=1}^{k} d^{-i}) l_0$. Substituting in the value of l_1 from Eq.(4.2), $\log n = (1 + \sum_{i=1}^{k+1} d^{-i}) \log n_0 = 1/\alpha \log n_0$. The number of points needed is $F(n,\epsilon) = \frac{k+2}{\epsilon} n_0^d = \frac{k+2}{\epsilon} n^{d\alpha}$.

How large is $d\alpha$? Consider the case when $d > 1$:

$$\frac{d}{(1 + \sum_{i=1}^{k+1} d^{-i})} = \frac{d-1}{1 - 1/d^{k+2}} = \frac{d^{k+2}}{d^{k+2}-1}(d-1)$$

Setting $k = \Omega(\log_d \log n)$, $F(n,\epsilon) = O(\frac{1}{\epsilon} n^{d-1} \log_d \log n)$. For example, $F(n,\epsilon) = O(\frac{1}{\epsilon} n \log\log n)$ when $d = 2$. Similarly, setting $k = \Omega(\log_{1/d} \log n)$ when $0 < d < 1$ yields $F(n,\epsilon) = O(\frac{1}{\epsilon} \log_{1/d} \log n)$.

When $d = 1$, $\frac{d}{(1+\sum_{i=1}^{k+1} d^{-i})} = \frac{1}{k+2}$. Setting $k = \Omega(\log n)$, we get $F(n,\epsilon) = O(\frac{1}{\epsilon} \log n)$.

5 Algorithms for Combinatorial Discrepancy

LEMMA 5.1. ([5]) *Combinatorial discrepancy for a set of red and blue points in the plane can be computed in time $O(n^2 \log n)$.*

Proof Sketch. [See [5] for details] A discrepancy maximizing rectangle has minimal and maximal points in the x and y dimensions. These four points fully describe the rectangle. By specifying the minimizing and maximizing y coordinates, the problem is reduced to a one-dimensional problem of all points within the slab this defines. By maintaining the maximal discrepancy interval in the one-dimensional problem under point insertion, only $O(n^2)$ insertions are necessary to check the maximum discrepancy interval over all possible slabs, and thus over all possible rectangles.

The one-dimensional problem is solved by building a binary tree of intervals. A node corresponding to interval I stores the subinterval i of maximal discrepancy, the interval l of maximal discrepancy that

includes the left boundary of I, and the interval r of maximal discrepancy that includes the right boundary of I. Two adjacent nodes, $I_{left} : (i_{left}, l_{left}, r_{left})$ and $I_{right} : (i_{right}, l_{right}, r_{right})$, can be merged to form a single nodes $I : (i, l, r)$ in constant time. i is assigned the interval with the maximum discrepancy out of the set $\{i_{left}, i_{right}, r_{left} \cup r_{left}\}$. l is assigned the interval with the maximum discrepancy out of the set $\{l_{left}, I_{left} \cup l_{right}\}$, and r is assigned symmetrically to l. The entire interval, $[0, 1] = I : (i, l, r)$, is at the root of the tree, and the interval which maximizes the discrepancy is i. Adding a point requires merging $O(\log n)$ intervals if the tree is balanced, and the tree can be kept balanced through rotations which only require a constant number of merge operations each.

THEOREM 5.1. *Let $\mathcal{R}'$ be the set of all rectangles such that $\sum_{p \in R} m(p)$ and $\sum_{p \in R} b(p)$ are greater than the constant C. Then any linear discrepancy function of the form $a_1 \sum m(p) + a_2 \sum b(p) + a_3$ can be maximized over this set in $O(n^2 \log n)$ time.*

Proof. Because a_3 is constant for all intervals, it can be ignored. Thus the linear function has the form of $d(m, b, R) = \sum_{p \in R} \chi(p)$. The algorithm of [5] relies only on the fact that the discrepancy function is additive, and hence can be extended to the above discrepancy function by only modifying the intervals and merge operation in the one-dimensional case.

Define $\mathcal{I}'$ to be the set of all intervals such that $\sum_{p \in I} m(p) \geq C$ and $\sum_{p \in I} b(p) \geq C$. For each interval $I : (i, l, r)$, i must be in $\mathcal{I}'$ and l and r must represent sets of intervals $l_1 \ldots l_k$ and $r_1 \ldots r_h$, respectively. l_k (resp. r_h) is the interval in $\mathcal{I}'$ which contains the left (resp. right) boundary that has the maximum discrepancy. l_1 (resp. r_1) is the interval which contains the left (resp. right) boundary that has the maximum discrepancy. For all i l_i includes the left boundary and $|l_i| < |l_j|$ for all $i < j$. Also $\sum_{p \in l_i} m(p) < C$ or $\sum_{p \in l_i} b(p) < C$ for all $i < k$. Finally, the set l must contain all local maximum; if l_i were to gain on lose one point, the discrepancy would decrease. The restrictions are the same for all r_i, expect these intervals must contain the right boundary.

The local optimality restriction ensures $\sum_{p \in l_i} m(p) < \sum_{p \in l_{i+1}} m(p)$ and $\sum_{p \in l_i} b(p) < \sum_{p \in l_{i+1}} b(p)$. If either measure ($\sum m(p)$ or $\sum b(p)$) increases then the other must also increase or this would violate the local optimality condition. An increase of just a measure that increases discrepancy will cause the previous interval not to be optimal and an increase in just a measure that causes the discrepancy to decrease will cause the latter interval not to be optimal. Thus k and h are constants bounded by the number of p required for $\sum m(p) \geq C$ and $\sum b(p) \geq C$. Thus each interval in the tree structure stores a constant amount of information.

A merge between two adjacent intervals $I_{left} : (i_{left}, l_{left}, r_{left})$ and $I_{right} : (i_{right}, l_{right}, r_{right})$ also can be done in constant time. Computing the maximum discrepancy interval in $\mathcal{I}'$ can be done by checking i_{left} and i_{right} versus all pairs from $l_{right} \cup r_{left}$ such that $\sum m(p) \geq C$ and $\sum b(p) \geq C$. There are a constant number of these. By the local optimality restriction, the optimal interval in $\mathcal{I}'$ spanning the boundary must have one endpoint in each set. Updating l and r can be done by just pruning from l_{left} and $I_{left} \cup r_{left}$ according to the restrictions for l, and similarly for r. These remain a constant size after the pruning. Because a merge can be done in constant time, a point can be added to the balanced tree in $O(\log n)$ time. Hence, the maximum discrepancy rectangle in $\mathcal{R}'$ for any linear discrepancy function can be found in $O(n^2 \log n)$ time.

A similar argument applies if we consider prospective discrepancy, or higher dimensions. We omit details.

LEMMA 5.2. *A linear discrepancy function can be maximized over prospective rectangles in $O(n^3 \log n)$ time, or it can be maximized over axis-parallel hyper-rectangles in d-dimensions in time $O(n^{2d-2} \log n)$.*

6 One-parameter Exponential Families

Having developed general algorithms for dealing with convex discrepancy functions, we now present a general expression for a likelihood-based discrepancy measure for the one-parameter exponential family. Many common distributions like the Poisson, Bernoulli, Gaussian and gamma distributions are members of this family. Subsequently we will derive specific expressions for the above mentioned distribution families.

DEFINITION 6.1. (ONE-PARAMETER EXP. FAMILY) *The distribution of a random variable y belongs to a one-parameter exponential family (denoted by $y \sim 1EXP(\eta, \phi, T, B_e, a)$ if it has probability density given by*

$$f(y; \eta) = C(y, \phi) exp((\eta T(y) - B_e(\eta))/a(\phi))$$

where $T(\cdot)$ is some measurable function, $a(\phi)$ is a function of some known scale parameter $\phi(> 0)$, η is an unknown parameter (called the natural parameter), and $B_e(\cdot)$ is a strictly convex function. The support $\{y : f(y; \eta) > 0\}$ is independent of η.

It can be shown that $E_\eta(T(Y)) = B_e^{'}(\eta)$ and $\mathrm{Var}_\eta(T(Y)) = a(\phi) B_e^{''}(\eta)$. In general, $a(\phi) \propto \phi$.

Let $\mathbf{y} = \{y_i : i \in R\}$ denote a set of $|R|$ variables that are independently distributed with $y_i \sim$ 1EXP(η, ϕ_i, T, b, a), $(i \in R)$. The joint distribution of $\mathbf{y}$ is given by

$$f(\mathbf{y};\eta) = \prod_{i \in R} C(y_i, \phi_i) exp((\eta T^*(\mathbf{y}) - B_e(\eta))/\phi^*)$$

where $1/\phi^* = \sum_{i \in R}(1/a(\phi_i))$, $v_i = \phi^*/a(\phi_i)$, and $T^*(\mathbf{y}) = \sum_{i \in R}(v_i T(y_i))$.

Given data $\mathbf{y}$, the *likelihood* of parameter η is the probability of seeing $\mathbf{y}$ if drawn from a distrbution with parameter η. This function is commonly expressed in terms of its logarithm, the *log-likelihood* $l(\eta;\mathbf{y})$, which is given by (ignoring constants that do not depend on η)

$$l(\eta;\mathbf{y}) = (\eta T^*(\mathbf{y}) - B_e(\eta))/\phi^* \tag{6.3}$$

and depends on data only through T^* and ϕ^*.

THEOREM 6.1. *Let $\mathbf{y} = (y_i : i \in R)$ be independently distributed with $y_i \sim$ 1EXP(η, ϕ_i, T, b, a), $(i \in R)$. Then, the maximum likelihood estimate (MLE) of η is $\hat{\eta} = g_e(T^*(\mathbf{y}))$, where $g_e = (B_e^{'})^{-1}$. The maximized log-likelihood (ignoring additive constants) is $l(\hat{\eta};\mathbf{y}) = (T^*(\mathbf{y}) g_e(T^*(\mathbf{y})) - B_e(g_e(T^*(\mathbf{y}))))/\phi^*$.*

Proof. The MLE is obtained by maximizing (6.3) as a function of η. Since B_e is strictly convex, $B_e^{'}$ is strictly monotone and hence invertible. Thus, the MLE obtained as a solution of $l(\eta;\mathbf{y})^{'} = 0$ is $\hat{\eta} = (B_e^{'})^{-1}(T^*(\mathbf{y})) = g_e(T^*(\mathbf{y}))$. The second part is obtained by substituting $\hat{\eta}$ in (6.3).

The likelihood ratio test for outlier detection is based on the following premise. Assume that data is drawn from a one-parameter exponential family. For a given region R_1 and its complement R_2, let η_{R_1} and η_{R_2} be the MLE parameters for the data in the regions. Consider the two hypothesis $H_0 : \eta_{R_1} = \eta_{R_2}$ and $H_1 : \eta_{R_1} \neq \eta_{R_2}$. The test then measures the ratio of the likelihood of H_1 versus the likelihood of H_0. The resulting quantity is a measure of the strength of H_1; the larger this number is, the more likely it is that H_1 is true and that the region represents a true outlier. The likelihood ratio test is *individually* the test with most statistical power to detect the region of maximum discrepancy and hence is optimal for the problems we consider. A proof of this fact for Poisson distributions is provided by Kulldorff [13] and extends to 1EXP without modification.

THEOREM 6.2. *Let $\mathbf{y_{R_j}} = (y_{R_j i} : i \in R_j)$ be independently distributed with $y_{R_j i} \sim$ 1EXP$(\eta_{R_j}, \phi_{R_j i}, T, B_e, a)$, for $j = 1, 2$. The log-likelihood ratio test statistic for testing $H_0 : \eta_{R_1} = \eta_{R_2}$ versus $H_1 : \eta_{R_1} \neq \eta_{R_2}$ is given by*

$$\Delta = \kappa(G_{R_1}, \Phi_{R_1}) + \kappa(G_{R_2}, \Phi_{R_2}) - \kappa(G, \Phi) \tag{6.4}$$

where $\kappa(x, y) = (x g_e(x) - B_e(g_e(x)))/y$, $G_{R_j} = T^*(\mathbf{y_{R_j}})$, $1/\Phi_{R_j} = \sum_{i \in R_j}(1/a(\phi_{R_j i}))$, $1/\Phi = 1/\Phi_{R_1} + 1/\Phi_{R_2}$, $b_{R_1} = \frac{1/\Phi_{R_1}}{(1/\Phi_{R_1} + 1/\Phi_{R_2})}$ and $G = b_{R_1} G_{R_1} + (1 - b_{R_1}) G_{R_2}$.

Proof. The likelihood ratio is given by $\frac{sup_{\eta_{R_1} \neq \eta_{R_2}} L(\eta_{R_1}, \eta_{R_2}; \mathbf{y_{R_1}}, \mathbf{y_{R_2}})}{sup_{\eta} L(\eta; \mathbf{y_{R_1}}, \mathbf{y_{R_2}})}$. Substituting the MLE expressions $\eta_{\hat{R}_1}$ and $\eta_{\hat{R}_2}$ from Theorem 6.1, and setting $G = T^*(\mathbf{y_{R_1}}, \mathbf{y_{R_2}}) = \frac{\sum_{j=1,2} \sum_{i \in R_j} T(y_{R_j i})/a(\phi_{R_j i})}{\sum_{j=1,2} \sum_{i \in R_j} 1/a(\phi_{R_j i})} = \frac{1/\Phi_{R_1}}{(1/\Phi_{R_1} + 1/\Phi_{R_2})} G_{R_1} + \frac{1/\Phi_{R_2}}{(1/\Phi_{R_1} + 1/\Phi_{R_2})} G_{R_2} = b_{R_1} G_{R_1} + (1 - b_{R_1}) G_{R_2}$, the result follows by computing logarithms.

FACT 6.1. *To test $H_0 : \eta_{R_1} = \eta_{R_2}$ versus $H_1 : \eta_{R_1} > \eta_{R_2}$, the log-likelihood ratio test statistic is given by*

$$\Delta = 1(\eta_{\hat{R}_1} > \eta_{\hat{R}_2})(\kappa(G_{R_1}, \Phi_{R_1}) + \kappa(G_{R_2}, \Phi_{R_2}) - \kappa(G, \Phi)) \tag{6.5}$$

Similar result holds for the alternative $H_1 : \eta_{R_1} < \eta_{R_2}$ with the inequalities reversed.

In the above expression for Δ (with $R_1 = R, R_2 = R^c$), the key terms are the values b_R and G_R. $G_R = T^*(\mathbf{y_R})$ is a function of the data (T^* is a *sufficient statistic* for the distribution), and thus is the equivalent of a measurement. In fact, G_R is a weighted average of $T(y_i)$s in R. Thus, $G_R/\Phi_R = \sum_{i \in R} T(y_i)/a(\phi_i)$ represents the *total* in R. Similarly, G/Φ gives the aggregate for the region and hence $m_R = \frac{\Phi}{\Phi_R}\frac{G_R}{G}$ is the fraction of total contained in R. Also, $1/\Phi_R$ gives the total area of R which is independent of the actual measurements and only depends on some baseline measure. Hence, $b_R = \frac{\Phi}{\Phi_R}$ gives the fraction of total area in R. The next theorem provides an expression for Δ in terms of m_R and b_R.

THEOREM 6.3. *Let $R_1 = R$ and $R_2 = R^c$. To obtain $R \in \mathcal{R}$ that maximizes discrepancy, assume G and Φ to be fixed and consider the parametrization of Δ in terms of b_R and $m_R = b_R G_R/G$.*

The discrepancy measure (ignoring additive constants) $d(.,.)$ is given by

$$\begin{aligned} d(m_R, b_R)\frac{\Phi}{G} &= m_R g_e(G\frac{m_R}{b_R}) - \frac{b_R}{G} B_e(g_e(G\frac{m_R}{b_R})) + \\ &\quad (1 - m_R) g_e(G\frac{1 - m_R}{1 - b_R}) - \\ &\quad \frac{(1 - b_R)}{G} B_e(g_e(G\frac{1 - m_R}{1 - b_R})) \end{aligned} \tag{6.6}$$

Proof. Follows by substituting $G_R = G\frac{m_R}{b_R}$, $G_{R^c} = G\frac{1-m_R}{1-b_R}$ in (6.4), simplifying and ignoring additive constants.

7 Discrepancy Measures For Specific Distributions

We can now put together all the results from the previous sections. Section 4 showed how to map a convex discrepancy function to a collection of linear discrepancy functions, and Section 5 presented algorithms maximizing general linear discrepancy functions over axis parallel rectangles. The previous section presented a general formula for discrepancy in a one-parameter exponential family. We will now use all these results to derive discrepancy functions for specific distribution families and compute maximum discrepancy rectangles with respect to them.

7.1 The Kulldorff Scan Statistic (Poisson distribution) The Kulldorff scan statistic was designed for data generated by an underlying Poisson distribution. We reproduce Kulldorff's derivation of the likelihood ratio test, starting from our general discrepancy function Δ.

In a Poisson distribution, underlying points are marked for the presence of some rare event (e.g. presence of some rare disease in an individual) and hence the measurement attached to each point is binary with a 1 indicating presence of the event. The number of points that get marked on a region R follows a Poisson process with base measure b and intensity λ if (i) $N(\emptyset) = 0$, (ii) $N(A) \sim \text{Poisson}(\lambda b(A)), A \subset R, b(\cdot)$ is a baseline measure defined on R and λ is a fixed intensity parameter (examples of $b(A)$ include the area of A, total number of points in A, *etc.*), and (iii) the number of marked points in disjoint subsets are independently distributed.

Derivation of the Discrepancy Function. A random variable $y \sim \text{Poisson}(\lambda\mu)$ is a member of 1EXP with $T(y) = y/\mu, \phi = 1/\mu, a(\phi) = \phi, \eta = \log(\lambda), B_e(\eta) = \exp(\eta), g_e(x) = \log(x)$. For a set of n independent measurements with mean $\lambda\mu_i, i = 1, \cdots, n$, $T^*(\mathbf{y}) = \sum_{i=1}^n y_i / \sum_{i=1}^n \mu_i, \phi^* = (\sum_{i=1}^n \mu_i)^{-1}$. For a subset R, assume the number of marked points follows a Poisson process with base measure $b(\cdot)$ and log-intensity η_R while that in R^c has the same base measure but log-intensity η_{R^c}. For any partition $\{A_i\}$ of R and $\{B_j\}$ of R^c, $\{N(A_i)\}$ and $\{N(B_j)\}$ are independently distributed Poisson variables with mean $\{\exp(\eta_R)b(A_i)\}$ and $\{\exp(\eta_{R^c})b(B_j)\}$ respectively. Then, $1/\Phi_R = \sum_{A_i} b(A_i)) = b(R)$, $1/\Phi_{R^c} = b(R^c)$, $G_R = \frac{\sum_{A_i} N(A_i)}{\sum_{A_i} b(A_i)} = N(R)/b(R)$, $G_{R^c} = N(R^c)/b(R^c)$, and $G = \frac{N(R)+N(R^c)}{b(R)+b(R^c)}$. Hence, $b_R = \frac{b(R)}{b(R)+b(R^c)}$ and $m_R = \frac{N(R)}{N(R)+N(R^c)}$.

$$\begin{aligned} d_K(b_R, m_R)\frac{\Phi}{G} &= m_R(\log(G) + \log(\frac{m_R}{b_R})) - b_R\frac{m_R}{b_R} + \\ &\quad (1-m_R)(\log(G) + \log(\frac{1-m_R}{1-b_R})) - \\ &\quad \frac{1-m_R}{1-b_R}(1-b_R) \\ &= m_R\log(\frac{m_R}{b_R}) + \\ &\quad (1-m_R)\log(\frac{1-m_R}{1-b_R}) + const \end{aligned}$$

and hence $d_K(b_R, m_R) = c(m_R\log(\frac{m_R}{b_R}) + (1-m_R)\log(\frac{1-m_R}{1-b_R}))$, where $c > 0$ is a fixed constant. Note that the discrepancy is independent of the partition used and hence is well defined.

Maximizing the Kulldorff Scan Statistic. It is easy to see that d_K is a convex function of m_R and b_R, is always positive, and grows without bound as either of m_R and b_R tends to zero. It is zero when $m_R = b_R$. The Kulldorff scan statistic can also be viewed as the Kullback-Leibler distance between the two two-point distributions $[m_R, 1-m_R]$ and $[b_R, 1-b_R]$. As usual, we will consider maximizing the Kulldorff scan statistic over the region $S_n = [1/n, 1-1/n]^2$. To estimate the size of an ϵ-approximate family for d_K, we will compute λ^* over S_n.

Let $f_K(x, y) = x\ln\frac{x}{y} + (1-x)\ln\frac{1-x}{1-y}$.

$$\nabla f_K = \mathbf{i}\left(\ln\frac{x}{1-x} + \ln\frac{y}{1-y}\right) + \mathbf{j}\left(\frac{x}{y} - \frac{1-x}{1-y}\right)$$

$$H(f_K) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{x(1-x)} & \frac{1}{y(1-y)} \\ \frac{1}{y(1-y)} & \frac{-x}{y^2} - \frac{1-x}{(1-y)^2} \end{pmatrix}$$

The eigenvalues of $H(f)$ are the roots of the equation $|H(f) - \lambda\mathbf{I}| = 0$. Solving for λ^*, and substituting from the expressions for the partial derivatives, and maximizing over S_n, we obtain $\lambda^* = \Theta(n)$.

Invoking Theorem 4.1 and Theorem 5.1,

THEOREM 7.1. *An additive ϵ-approximation to the maximum discrepancy d_K over all rectangles containing at least a constant measure can be computed in time $O(\frac{1}{\epsilon}n^2\log^2 n)$. With respect to prospective time windows, the corresponding maximization takes time $O(\frac{1}{\epsilon}n^3\log^2 n)$.*

The Jensen-Shannon divergence is a symmetrized variant of the Kullback-Leibler distance. We mentioned earlier that d_K can be expressed as a Kullback-Leibler

distance. Replacing this by the Jensen-Shannon distance, we get a symmetric version of the Kulldorff statistic, for which all the bounds of Theorem 7.1 apply directly.

7.2 Gaussian Scan Statistic It is more natural to use an underlying Gaussian process when measurements are real numbers, instead of binary events. In this section, we derive a discrepancy function for an underlying Gaussian process. To the best of our knowledge, this derivation is novel.

Derivation of the Discrepancy Function. A random variable y that follows a Gaussian distribution with mean μ and variance $1/\tau^2$ (denoted as $y \sim N(\mu, 1/\tau^2)$ is a member of 1EXP with $T(y) = y, \eta = \mu, B_e(\eta) = \eta^2/2, \phi = 1/\tau^2, a(\phi) = \phi, g_e(x) = x$. For a set of n independent measurements with mean μ and variances $1/\tau_i^2, i = 1, \cdots, n$(known), $\phi^* = (\sum_{i=1}^n \tau_i^2)^{-1}$ and $T^*(\mathbf{y}) = \sum_{i=1}^n y_i\tau_i^2 / \sum_{i=1}^n \tau_i^2$. Assume measurements in R are independent $N(\mu_R, 1/\tau_i^2), (i \in R)$ while those in R^c are independent $N(\mu_{R^c}, 1/\tau_i^2), (i \in R^c)$. Then, $\Phi_R = (\sum_{i\in R} \tau_i^2)^{-1}$,$\Phi_{R^c} = (\sum_{i\in R^c} \tau_i^2)^{-1}$, $G_R = \frac{\sum_{i\in R} \tau_i^2 y_i}{\sum_{i\in R} \tau_i^2}$, $G_{R^c} = \frac{\sum_{i\in R^c} \tau_i^2 y_i}{\sum_{i\in R^c} \tau_i^2}$, and $G = \frac{\sum_{i\in R+R^c} \tau_i^2 y_i}{\sum_{i\in R+R^c} \tau_i^2}$. Hence, $b_R = \frac{1/\Phi_R}{(1/\Phi_R + 1/\Phi_{R^c})} = \frac{\sum_{i\in R} \tau_i^2}{\sum_{i\in R+R^c} \tau_i^2}$ and $m_R = \frac{\sum_{i\in R} \tau_i^2 y_i}{\sum_{i\in R+R^c} \tau_i^2}$. Thus,

$$\begin{aligned} d_G(b_R, m_R)\frac{\Phi}{G} &= m_R G\frac{m_R}{b_R} - \frac{b_R}{G}G\frac{m_R}{b_R} + \\ &\quad (1-m_R)G\frac{1-m_R}{1-b_R} - \frac{1-b_R}{G}G\frac{1-m_R}{1-b_R} \\ &= G(\frac{m_R^2}{b_R} + \frac{(1-m_R)^2}{1-b_R}) - 1 = G\frac{(m_R-b_R)^2}{b_R(1-b_R)} \end{aligned}$$

and hence $d_G(b_R, m_R) = c\frac{(m_R-b_R)^2}{b_R(1-b_R)}$, where $c > 0$ is a fixed constant. Note that the underlying baseline $b(\cdot)$ is a weighted counting measure which aggregate weights τ_i^2 attached to points in a region.

Maximizing the Gaussian Scan Statistic. Again, it can be shown that d_G is a convex function of both parameters, and grows without bound as b_R tends to zero or one. Note that this expression can be viewed as the χ^2-distance between the two two-point distributions $[m_R, 1-m_R], [b_R, 1-b_R]$. The complexity of an ϵ-approximate family for d_G can be analyzed as in Section 7.1. Let $f_G(x, y) = \frac{(x-y)^2}{y(1-y)}$. See full version of paper for expressions for ∇f_G and $H(f_G)$. Solving the equation $|H - \lambda\mathbf{I}|$, and maximizing over S_n, we get $\lambda^* = O(n^2)$.

THEOREM 7.2. *An additive ϵ-approximation to the maximum discrepancy d_G over all rectangles containing at least a constant measure can be computed in time $O(\frac{1}{\epsilon}n^3 \log n \log\log n)$. With respect to prospective time windows, the corresponding maximization takes time $O(\frac{1}{\epsilon}n^4 \log n \log\log n)$.*

Trading Error for Speed For the Kulldorff statistic, the function value grows slowly as it approaches the boundaries of S_n. Thus, only minor improvements can be made when considering relative error approximations. However, for the Gaussian scan statistic, one can do better. A simple substitution shows that when $x = 1 - \frac{1}{n}$, $y = \frac{1}{n}$, $f_G(x, y) = \Theta(n)$. Using this bound in Theorem 4.1, we see that a relative ϵ-approximate family of size $O(\frac{1}{\epsilon}\log n)$ can be constructed for d_G, thus yielding the following result:

THEOREM 7.3. *A $1/(1+\epsilon)$ approximation to the maximum discrepancy d_G over the space of axis parallel rectangles containing constant measure can be computed in time $O(\frac{1}{\epsilon}n^2 \log^2 n)$.*

7.3 Bernoulli Scan Statistic Modeling a system with an underlying Bernoulli distribution is appropriate when the events are binary, but more common than those that would be modeled with a Poisson distribution. For instance, a baseball player's batting average may describe a Bernoulli distribution of the expectation of a hit, assuming each at-bat is independent.

Derivation of the Discrepancy Function. A binary measurment y at a point has a Bernoulli distribution with parameter θ if $P(y = 1) = \theta^y(1-\theta)^{1-y}$. This is a member of 1EXP with $T(y) = y, \eta = \log(\frac{\theta}{1-\theta}), B_e(\eta) = \log(1+\exp(\eta)), \phi = 1, a(\phi) = 1, g_e(x) = \log(x) - \log(1-x)$.

For a set of n independent measurements with parameter η, $\phi^* = 1/n, T^*(\mathbf{y}) = \sum_{i=1}^n y_i/n$. Assuming measurements in R and R^c are independent Bernoulli with parameters η_R and η_{R^c} respectively, $\Phi_R = 1/|R|, \Phi_{R^c} = 1/|R^c|, G_R = y(R)/|R|, G_{R^c} = y(R^c)/|R^c|, b_R = \frac{|R|}{|R|+|R^c|}, G = \frac{y(R)+y(R^c)}{|R|+|R^c|}, m_R = \frac{y(R)}{y(R)+y(R^c)}$. Note that $y(A)$ denotes the number of 1's in a subset A. Thus,

$$\begin{aligned} d_B(b_R, m_R)\frac{\Phi}{G} &= m_R \log(\frac{m_R}{b_R}) + \\ &\quad (1-m_R)\log(\frac{1-m_R}{1-b_R}) + (\frac{b_R}{G} - m_R)\log(1 - G\frac{m_R}{b_R}) \\ &\quad + (\frac{1-b_R}{G} - 1 + m_R)\log(1 - G\frac{1-m_R}{1-b_R}) \end{aligned}$$

Maximizing the Bernoulli Scan Statistic. Much like d_K, it is easy to see that d_B is a convex function of m_R and b_R, is always positive, and grows without bound as either b_R or m_R tend to zero

or one. The complexity of an ϵ-approximate family for d_B, the Bernoulli scan statistic, can be analyzed by letting $f_B(x,y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y} + \left(\frac{y}{G} - x\right)\log\left(1 - G\frac{x}{y}\right) + \left(\frac{1-y}{G} - 1 + x\right)\log\left(1 - G\frac{1-x}{1-y}\right)$, where G is a constant. See full version of the paper for expressions for ∇f_B and $H(f_B)$. Direct substitution of the parameters yields $\lambda^* = O(n)$.

THEOREM 7.4. *An additive ϵ-approximation to the maximum discrepancy d_B over all rectangles containing at least a constant measure can be computed in time $O(\frac{1}{\epsilon}n^2\log^2 n)$. With respect to prospective time windows, the corresponding maximization takes time $O(\frac{1}{\epsilon}n^3\log^2 n)$.*

7.4 Gamma Scan Statistic When events arrive one after another, where a Poisson variable describes the interval between events, then a gamma distribution describes the count of events after a set time.

Derivation of the Discrepancy Function. A positive measurement y has a gamma distribution with mean $\mu(>0)$ and shape $\nu(>0)$ if it has density $\frac{\nu^\nu}{\mu^\nu\Gamma(\nu)}\exp(-\frac{\nu}{\mu}y)x^{\nu-1}$ and is a member of 1EXP with $T(y) = y, \eta = -\frac{1}{\mu}(<0), B_e(\eta) = -\log(-\eta), \phi = 1/\nu, a(\phi) = \phi, g_e(x) = -\frac{1}{x}$. Following arguments similar to the Gaussian case, $\Phi_R = (\sum_{i\in R}\nu_i)^{-1}, \Phi_{R^c} = (\sum_{i\in R^c}\nu_i)^{-1}, G_R = \frac{\sum_{i\in R}\nu_i y_i}{\sum_{i\in R}\nu_i}, G_{R^c} = \frac{\sum_{i\in R^c}\nu_i y_i}{\sum_{i\in R^c}\nu_i}, G = \frac{\sum_{i\in R+R^c}\nu_i y_i}{\sum_{i\in R+R^c}\nu_i}$. Hence, $b_R = \frac{1/\Phi_R}{(1/\Phi_R + 1/\Phi_{R^c})} = \frac{\sum_{i\in R}\nu_i}{\sum_{i\in R+R^c}\nu_i}$ and $m_R = \frac{\sum_{i\in R}\nu_i y_i}{\sum_{i\in R+R^c}\nu_i y_i}$. Thus,

$$\begin{aligned} d_\gamma(b_R, m_R)\frac{\Phi}{G} &= m_R(-\frac{b_R}{Gm_R}) - \frac{b_R}{G}\log(G\frac{m_R}{b_R}) + \\ &\quad (1-m_R)(-\frac{1-b_R}{G(1-m_R)}) - \frac{1-b_R}{G}\log(G\frac{1-m_R}{1-b_R}) \\ &= b_R\log(\frac{b_R(1-m_R)}{m_R(1-b_R)}) - \log(\frac{1-m_R}{1-b_R}) + const \\ &= b_R\log(\frac{b_R}{m_R}) + (1-b_R)\log(\frac{1-b_R}{1-m_R}) + const \end{aligned}$$

and hence ignoring additive constants, $d_\gamma(b_R, m_R) = c(b_R\log(\frac{b_R}{m_R}) + (1-b_R)\log(\frac{1-b_R}{1-m_R})), c(>0)$ is a fixed constant. For a fixed shape parameter (i.e. $\nu_i = \nu$ for each i), $b_R = \frac{|R|}{|R|+|R^c|}$ and $m_R = \frac{\sum_{i\in R} y_i}{\sum_{i\in R+R^c} y_i}$.

Maximizing the Gamma Scan Statistic. Because $d_\gamma = d_K$ up to an additive constant, $f_\gamma = f_K$ and thus $\lambda^* = O(n)$ for $H(f_\gamma)$.

THEOREM 7.5. *An additive ϵ-approximation to the maximum discrepancy d_γ over all rectangles containing at least a constant measure can be computed in time $O(\frac{1}{\epsilon}n^2\log^2 n)$. With respect to prospective time windows, the corresponding maximization takes time $O(\frac{1}{\epsilon}n^3\log^2 n)$.*

References

[1] CORMODE, G., KORN, F., MUTHUKRISHNAN, S., AND SRIVASTAVA, D. Diamond in the rough: finding hierarchical heavy hitters in multi-dimensional data. In *SIGMOD* (2004), pp. 155–166.

[2] CORMODE, G., AND MUTHUKRISHNAN, S. What's hot and what's not: tracking most frequent items dynamically. In *PODS* (2003), pp. 296–306.

[3] CRESSIE, N. *Statistics for spatial data*, 2nd ed. John Wiley, 1993.

[4] DOBKIN, D., AND EPPSTEIN, D. Computing the discrepancy. *Symposium on Computational Geometry 9* (1993).

[5] DOBKIN, D. P., GUNOPULOS, D., AND MAASS, W. Computing the maximum bichromatic discrepancy, with applications to computer graphics and machine learning. *NeuroCOLT Technical Report Series NC-TR-95-008* (March 1995).

[6] FRIEDMAN, J. H., AND FISHER, N. I. Bump hunting in high-dimensional data. *Statistics and Computing 9*, 2 (April 1999), 123–143.

[7] GOOD, P. *Permutation tests - a practical guide to resampling methods for testing hypotheses*. Springer-Verlag, 2nd edition, New York, 2000.

[8] IYENGAR, V. On detecting space-time clusters. *KDD* (2004), 587–592.

[9] J.GLAZ, AND BALAKRISHNAN, N. *Scan statistics and applications*. Birkhauser, Boston, 1999.

[10] J.GLAZ, J.NAUS, AND S.WALLENSTEIN. *Scan statistics*. Springer-Verlag,New York, 2001.

[11] LIPSON, D., AUMANN, Y., BEN-DOR, A., LINIAL, N., AND YAKHINI, Z. Efficient calculation of interval scores for dna copy number data analysis. *RECOMB* (2005), 83–100.

[12] M.DWASS. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics 28* (1957), 181–187.

[13] M.KULLDORFF. A spatial scan statistic. *Communications in Statistics: Theory and Methods 26* (1997), 1481–1496.

[14] M.KULLDORFF. Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal statistical society,series A 164* (2001), 61–72.

[15] NEILL, D. B., AND MOORE, A. W. A fast multi-resolution method for detection of significant spatial disease clusters. *Advances in Neural Information Processing Systems 10* (2004), 651–658.

[16] NEILL, D. B., AND MOORE, A. W. Rapid detection of significant spatial clusters. In *KDD* (2004).