The Hunting of the Bump: On Maximizing Statistical Discrepancy

Deepak Agarwal* Jeff M. Phillips† Suresh Venkatasubramanian‡

1 Introduction

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.

Consider the cardinality n point set P where each point $p \in P$ is applied to a *baseline function* $b : P \to \mathbb{R}$ (what we expect to observe) and to a *measurement function* $m : P \to \mathbb{R}$ (what we actually observe). For any range R in set of ranges $\mathcal{R}$ we can define a *discrepancy function* $d_R : (m, b) \to \mathbb{R}$, which measures how different the observed measurements m_R are from the expected measurements b_R within the range R. We provide efficient approximation algorithms, both additive and relative, to solve the following problem:

Problem 1.1. *Given a point set P with baseline and measurement functions m and b, a range space $X = (P, \mathcal{R})$ where $\mathcal{R}$ describes all axis-aligned rectangles, and a convex discrepancy function d, find the range $R \in \mathcal{R}$ that maximizes d.*

2 Prior Work

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 [2, Chapter 8]). The spatial scan statistic, an important example of a convex discrepancy function $d_K(b_R, m_R) = m_R \log \frac{m_R}{b_R} + (1 - m_R) \log \frac{1-m_R}{1-b_R}$, introduced by Kulldorff [6] provides an elegant solution for detection and evaluation of spatial clusters. The technique has found wide applicability in public health, biosurveillance, environmental monitoring *etc*.

A brute force technique can solve Problem 1.1 for points in the plane, for any discrepancy function in $O(n^4)$. A *linear discrepancy function* is defined $\Delta_R(m, b) = c_1 \sum m(p) + c_2 \sum b(p) + c_3$. Dobkin, Maass and Gunopoulos [3] solve Problem 1.1 for the specific linear discrepancy function known as *combinatorial discrepancy* where $c_1 = 1$, $c_2 = -1$ and $c_3 = 0$, in $O(n^2 \log n)$ time for points in the plane.

Other related algorithmic work is heuristic and makes no guarantee on the quality of the solution, such as work by Iyengar [5], and Friedman and Fisher [4], or is conservative in that it provides an exact solution which in practice runs fast but reverts to brute force in the worst case, such as work by Neill and Moore [8, 7]

3 Our Contribution

Our main result, see [1] for a full version, 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. We show that the Dobkin *et. al.* algorithm can be extended 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 for points in the plane; as an example, we present an additive approximation algorithm for the Kulldorff scan statistic that runs in time $O(\frac{1}{\epsilon} n^2 \log^2 n)$, whereas an exact, brute force approach runs in $O(n^4)$ time.

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 want to maximize a general discrepancy function over a general shape class, you need only consider linear 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 underly-

*AT&T Labs – Research dagarwal@research.att.com
†Duke University jeffp@cs.duke.edu
‡AT&T Labs – Research suresh@research.att.com

	Our results		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 higher dimensions d, multiply by n^{2d-4}.

ing 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.

4 Convex Approximation

We present a general approximation theorem for maximizing a convex discrepancy function d. First we rephrase problem 1.1 to the following equivalent problem.

Problem 4.1. *Maximize convex discrepancy function d over all points $r = (m_R, b_R), R \in \mathcal{R}$.*

Using this formulation we can describe how to approximate a convex discrepancy function with a family of linear discrepancy functions. Let $\ell(x, y) = c_1 x + c_2 y + c_3$ 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 $l^U(x, y) \le d(x, y) \le l^U(x, y) + \epsilon$

Next we link the approximation error to the Hessian of the discrepancy function.

Lemma 4.1. *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 $\mathbf{q}$.*

Let $\lambda^* = \sup_{\mathbf{p} \in S_n} \lambda_{\max}(H(f)(\mathbf{p}))$. Let $\epsilon_{\mathbf{p}}(\mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|^2 \lambda^*$. In the approximation we need to consider all possible $(m_R, b_R) \in S_n = [C/n, 1 - C/n]^2$, for constant C. This restriction requires that each range contains some minimum level of support.

Lemma 4.2. *Let $\mathcal{C} \subset S_n$ be a set of t points such that for all $\mathbf{q} \in S_n, \min_{\mathbf{p} \in \mathcal{C}} \epsilon_{\mathbf{p}}(\mathbf{q}) \le \epsilon$. Then the t tangent planes at the points $f(\mathbf{p}), \mathbf{p} \in \mathcal{C}$, form an ϵ-approximate family for f.*

Finally, our main theorem uses a stratified grid decomposition to utilize the dependence on the approximation error on the Hessian.

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. Let $\lambda(n) = O(n^c)$. Then,*

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

The maximum discrepancy point $r = (m_R, b_R)$ over all linear discrepancy functions is an ϵ-approximation for the convex discrepancy function. A relative approximation theorem is similar.

References

[1] AGARWAL, D., PHILLIPS, J. M., AND VENKATASUBRAMANIAN, S. The Hunting of the Bump: On Maximizing Statistical Discrepancy. *SODA* (2006).

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

[3] 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).

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

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

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

[7] 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.

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