Contamination Estimation via Convex Relaxations

Matthew L. Malloy
comScore
mmalloy@comscore.com

Scott Alfeld
University of Wisconsin
salfeld@cs.wisc.edu

Paul Barford
comScore, University of Wisconsin
pb@cs.wisc.edu

***Abstract*—Identifying anomalies and contamination in datasets is important in a wide variety of settings. In this paper, we describe a new technique for estimating contamination in large, discrete valued datasets. Our approach considers the normal condition of the data to be specified by a model consisting of a set of distributions. Our key contribution is in our approach to contamination estimation. Specifically, we develop a technique that identifies the minimum number of data points that must be discarded (*i.e.,* the level of contamination) from an empirical data set in order to match the model to within a specified *goodness-of-fit*, controlled by a p-value. Appealing to results from large deviations theory, we show a lower bound on the level of contamination is obtained by solving a series of convex programs. Theoretical results guarantee the bound converges at a rate of $O(\sqrt{\log(p)/p})$, where p is the size of the empirical data set.**

Index terms: contamination estimation, anomaly detection, entropy minimization, discrete goodness-of-fit testing.

I. Introduction

Anomalies in datasets are typically associated with unexpected or unwanted characteristics such as contamination, noise or outliers that deviate significantly from expectations. The ability to detect anomalies and accurately estimate contamination in datasets is important in a wide variety of domains including healthcare, astronomy, environmental and materials sciences. The context that motivates our work is detecting anomalies and estimating contamination in datasets collected from communication and computer systems. Specific applications of anomaly detection in these datasets include network management and Internet security broadly defined. Communication and Internet measurement datasets have several distinguishing characteristics including the potential for extreme scale and high dimensionality.

The standard framework for anomaly detection is based on establishing a baseline for *normal* (*e.g.,* in a distributional sense) and then setting a threshold which if exceeded identifies an anomaly. The goal in establishing norms and thresholds is to identify anomalies with low false alarm rates. There is an extensive literature on methods for anomaly detection (see related work in Section III).

In this paper we describe a new method for anomaly detection which is based on estimating the level of contamination in a dataset. An anomaly is declared if a dataset has an elevated level of contaminate. We consider the contamination-free (*i.e.,* normal) condition of a dataset to be specified by a model comprised of a set of distributions. We then compare the model to the distributional profile of a target dataset collected over a specified period. A standard method for comparing datasets in this way is goodness of fit (GoF) testing [1]. To the best of our knowledge, this paper is the first to address the problem of contamination estimation using GoF testing based on entropy minimization, as we define in Section II-B.

The approach we develop is based on answering the following question. Given a model consisting of a family of distributions, a specified p-value, and an empirical dataset, what is the minimum number of data points that must be discarded so that the empirical distribution of the data matches a member model distribution (in terms of GoF for a specified p-value)? This is akin to finding the largest subset of the original dataset which has an empirical distribution *close* to the model. We show that this question can be efficiently answered by solving a series of convex optimizations. Solving the optimizations results in a lower bound on the minimum number of data points that are attributed to a contaminate. In the simplest case, each convex optimization is an inequality constrained entropy minimization problem (whose dual is a constrained geometric program) which can be solved in real time and at scale for many applications. More generally, the approach can be applied to any setting in which the model consists of a convex set of distributions. Two specific instances which we discuss are *1)* models defined by any number of distributions with arbitrary mixture proportions, and *2)* models defined by the set distributions with small Kullback-Leibler (KL) divergence to a specified distribution, which arises when the model itself is generated from a finite amount of data. Lastly, we show the lower bound output by the optimization converges to an upper bound known as the separation distance at a rate of $O(\sqrt{\log(p)/p})$, where p is the number of data points.

II. Quantifying Contamination

A. Notation

Let $P \in \mathbb{R}^n$ and $Q \in \mathbb{R}^n$ denote probability mass functions over n categories, with elements P_i, $i = 1, \ldots, n$ and Q_i, $i = 1, \ldots, n$. Throughout, P denotes the distribution under test, Q denotes a member distribution of the model, Q^0 denotes the 'true' unknown model distribution, and Q^j indexes multiple distributions. The empirical distribution of a sequence of random variables $X = X_1, \ldots, X_p \in \mathcal{X}^p$ is the relative proportion of occurrences of each element of $\mathcal{X}$ in X. Specifically, let $\mathcal{X} =: \{x_1, x_2, \ldots, x_n\}$ and define $p_i = \sum_{j=1}^{p} \mathbf{1}_{\{X_j = x_i\}}$ for $i = 1, \ldots, n$. Then $\widehat{P}(X) = \frac{1}{p}\{p_1, p_2, \ldots, p_n\}$. $\mathbb{P}_Q(\cdot)$ denotes probability measure with respect to distribution

Q. For simplicity of notation, we write $\mathbb{P}_Q(\{\widehat{P}^1, \widehat{P}^2\})$ as short hand for $\mathbb{P}_Q\left(\left\{X \in \mathcal{X}^p : \widehat{P}(X) \in \{\widehat{P}^1, \widehat{P}^2\}\right\}\right)$. The Kullback-Leibler divergence between two distributions is defined in the usual manner,

$$D(P||Q) := \sum_i P_i \log\left(\frac{P_i}{Q_i}\right).$$

$D(P||Q)$ is a jointly convex function in P and Q. The minimum entropy set, $\{P : D(P||Q) \leq \epsilon\}$, is a convex set (for a fixed Q, ϵ). Lastly, let $\mathbb{S}^n$ denote the probability simplex:

$$\mathbb{S}^n := \left\{P \in \mathbb{R}^n : \sum_i P_i = 1,\ P_i \geq 0 \quad i = 1, \ldots, n\right\}.$$

B. Quantifying Contamination

Consider a set of model distributions $\mathcal{Q}$ whose elements are supported over a finite number of categories $\mathcal{X}$ with $|\mathcal{X}| = n$. For example, $\mathcal{Q}$ could be set of minimum entropy distributions, or a mixture distribution, $Q = \sum_{j=1}^{\ell} \pi_j Q^j$, where $\pi_1, \ldots, \pi_\ell$ are unknown ($\mathcal{Q}$ is the set of all such mixture distributions). Let $X \in \mathcal{X}^p$ denote a collection of samples. An unknown subset of the samples consists of $i.i.d.$ draws from an unknown distribution $Q \in \mathcal{Q}$. The remaining samples, $\mathcal{C} \subset [p]$, are generated by some other means, and correspond to *contaminated* samples. This paper is concerned with lower bounding the size of the contaminating set $\mathcal{C}$ given the set of model distributions $\mathcal{Q}$, a specified significance level (a p-value), and the observed samples $X_1, \ldots, X_p$.

Intuitively, if the empirical distribution of a sequence of random variables is *close* to the model distribution in terms of GoF, we conclude the sequence is *not* contaminated. To quantify this intuition, we define a set of *typical* empirical distributions based on statistical significance; we note this definition is distinct from the usual definitions of *strongly* and *weakly* typical, and making this connection is a contribution herein.

Definition 1. Typical. Let $\widehat{P}^1, \widehat{P}^2, \ldots$ be an ordering on all empirical distributions (of p samples and n categories) such that $\mathbb{P}_Q(\widehat{P}^1) \leq \mathbb{P}_Q(\widehat{P}^2) \leq \ldots$. A sequence of random variables X with $\widehat{P}(X) = \widehat{P}^\ell$ is *typical* at significance level ϵ with respect to $\mathcal{Q}$ iff

$$\sup_{Q \in \mathcal{Q}} \mathbb{P}_Q\left(\left\{\widehat{P}^1, \widehat{P}^2, \ldots, \widehat{P}^{\ell-1}, \widehat{P}^\ell\right\}\right) \geq \epsilon \tag{1}$$

for any such ordering[1].

The definition implies a sequence of random variables X is typical if the probability of the empirical distribution of X *or any less likely empirical distribution* is more than a specified significance level. Note ϵ is interpreted as a p-value; as ϵ approaches zero, all sequences become typical (requiring stronger evidence to reject the null hypothesis). As ϵ increases, fewer sequences are typical.

[1]Note the ordering is an implicit function of Q; we suppress this for simplicity of notation.

Definition 2. Contaminated. We say X is *contaminated* iff X is not typical (with respect to $\mathcal{Q}$ and with significance ϵ). Likewise, an empirical distribution $\widehat{P}(X)$ is *contaminated* iff X is not typical.

In this paper we study the following question. Let $X = X_1, \ldots, X_p$ be a dataset, and let $X_{\widehat{\mathcal{C}}} = \{X_i : i \in \widehat{\mathcal{C}}\}$ be any subset of of the original dataset. What is the smallest set $\widehat{\mathcal{C}} \subset [p]$ such that $X_{[p]\setminus\widehat{\mathcal{C}}}$ *is not contaminated*? Specifically, let

$$c^* = \inf\left\{|\widehat{\mathcal{C}}| : X_{[p]\setminus\widehat{\mathcal{C}}} \text{ is typical for } (\mathcal{Q}, \epsilon)\right\}.$$

How and under what conditions can one compute c^* efficiently? Our main focus and insight will be on the continuous approximation to c^*/p, denoted α^*:

$$\alpha^* = \inf\{\alpha \in [0,1] : \exists P \in \mathcal{P}(X, \alpha) \text{ typical for } (\mathcal{Q}, \epsilon)\}$$

where $\mathcal{P}(X, \alpha)$ is the set of all distributions that can be created by discarding a fraction α of the mass of $\widehat{P}(X)$ (see Sec. II-D):

$$\mathcal{P}(X, \alpha) = \left\{P \in \mathbb{S}^n : P_i \leq \frac{\widehat{P}_i(X)}{1-\alpha} \quad i = 1, \ldots, n\right\}. \tag{2}$$

Throughout, α is a key parameter that represents the fraction of the dataset attributed to contamination; α^* represents the smallest α such that there exists a subset of the original data of size $p(1-\alpha)$ that is *not* contaminated. If $\alpha^* = 0$, the original dataset is not contaminated; if $\alpha^* = 1$, the entire dataset must be attributed to contamination.

C. Separation Distance

We assume $X_i \overset{i.i.d.}{\sim} Q^0$ for all $i \notin \mathcal{C}$. For X_i, $i \in \mathcal{C}$, no assumption is made. This agnostic approach has inherent limitations. In the extreme case the distribution of the contaminated data could exactly follow that of the model. Here, the distribution of the full dataset should closely match the model, and be indistinguishable from the setting where $\mathcal{C}$ is empty. No contamination should be reported to within the significance level (in m realizations of X^p, we expect $c^* \neq 0$ fewer than $m\epsilon$ times).

A more interesting scenario is when the empirical distribution of the full dataset converges to a *distinct* distribution *i.e.*, $\widehat{P}(X^p) \to P \neq Q^0$. In the case that $\mathcal{Q} = \{Q^0\}$, a consistent estimator will report non-zero contamination for large p. P can be written as a mixture distribution, and we are interested in reporting the smallest κ such that $(1-\kappa)Q^0 + \kappa F = P$ for *any* distribution F. F represents the contaminating distribution, and κ the proportion of the samples which are drawn from F. This minimum value of κ is known as the *separation distance* [2] between P and Q^0, written succinctly as

$$\kappa(P||Q^0) = \max_{i \in [n]}\left(1 - \frac{P_i}{Q_i^0}\right).$$

In this way, the separation distance between the empirical distribution of the data and model distribution plays an important role in the behavior of c^* and α^* as the sample size grows. We show as a corollary to later results that α^* is both upper bounded by and converges to $\kappa(\widehat{P}(X)||Q^0)$ as p grows (see Proposition 1 and Theorem 6).

D. Convex Relaxations

With the exception of problems involving data over only two categories ($n = 2$), directly checking if a sample is contaminated is computationally prohibitive, even in the setting where the model consists of a single distribution (when $\mathcal{Q} = \{Q^0\}$). Alternatively, using large deviations results, bounds can be derived. The bound presented below can confirm if a particular dataset is contaminated. The theorem involves the KL divergence between the empirical distribution and a member of $\mathcal{Q}$. In the case where $\mathcal{Q} = \{Q^0\}$, the bound provides a simple way to check if a sample is contaminated at a particular significance level ϵ; in the more general case, if $\mathcal{Q}$ is a convex set, numerical optimization techniques can efficiently check the condition.

Theorem 1. *(Outer Bound). If*

$$\inf_{Q \in \mathcal{Q}} D(\widehat{P}(X)||Q) \geq \frac{1}{p} \log\left(\frac{1}{\epsilon}\right) + \frac{2n}{p} \log(p+1) \tag{3}$$

then X is contaminated at significance level ϵ.

Proof. See Appendix A. □

Theorem 1 is an outer bound; any empirical distribution with KL distance *greater* than the stated quantity (from *all* elements in $\mathcal{Q}$) *is* contaminated. Theorem 1 can be used to bound the size of the smallest set $\mathcal{C} \subset [p]$ such that $X_{[p]\setminus\mathcal{C}}$ *is not contaminated*. This is simplified if $\mathcal{Q}$ consists of a single model distribution; we first discuss this scenario. In principle, given a dataset $X \in \mathcal{X}^p$ and a model distribution Q^0, one could first check if X is contaminated by evaluating (3). If (3) holds, X is contaminated, and an immediate question follows – how many and which data points must be excluded so that (3) no longer holds? A exhaustive approach to answer this question would be the following. For each $x_i \in \mathcal{X}$, discard a single data point that takes the value x_i, and recalculate the empirical distribution with the data point removed. Of the n new empirical distributions, check if the one with minimum KL divergence to the model distribution still satisfies (3).

If (3) still holds for all possible empirical distributions with one data point removed, check all distinct empirical distributions that can be created by discarding 2 data points (roughly n^2 possibilities, provided each x_i appears at least twice in the data). Continuing in this manner, one would check each of the $\sim n^m$ possible empirical distributions that can be created by discarding m data points. When (3) is first violated, m lower bounds the minimum number of data points that must be excluded to match the model. We can interpret this as a series of integer programs. For $m = 0, \ldots, p$ define D_m^* as the solution to

$$\begin{aligned} \underset{m_1, m_2, \ldots, m_n \in \mathbb{N}^n}{\text{minimize}} \quad & \sum_{i=1}^{n} \frac{p_i - m_i}{p - m} \log\left(\frac{\frac{p_i - m_i}{p-m}}{Q_i^0}\right) \\ \text{subject to} \quad & \sum_i m_i = m \\ & m_i \leq p_i \qquad i = 1, \ldots, n \end{aligned} \tag{4}$$

where p_i is the number of times x_i appears in the original dataset X. The optimization variables, m_i, represent the number of samples to discard corresponding to a particular x_i. Note that the objective is the KL divergence between the *new* empirical distribution (with m samples removed) and the known distribution Q^0. The value of D_m^* can be checked in Theorem 1, providing conditions under which one can find a set $|\mathcal{C}| = m$ such that $X_{[p]\setminus\mathcal{C}}$ is not contaminated. This gives a bound on c^*. Specifically,

$$c^* \geq \max\left\{ m : D_m^* \geq \frac{1}{p-m} \log\left(\frac{1}{\epsilon}\right) + \frac{2n}{p-m} \log(p-m+1) \right\}.$$

Note that the condition in Theorem 1 will always be met for some m; in particular, for $m = p$, by convention $D_0^* = 0$, implying that the empty set, $X_{\{\}}$, is not contaminated.

The optimization in (4) is an integer program over a subset of $\mathbb{N}^n$. To efficiently solve the optimization, we can translate the integer valued variables to their continuous counterparts; specifically, let $\widehat{P}_i = p_i/p$, be the original empirical distribution, and $\alpha = m/p$ represent the fraction the total samples discarded. Making these substitutions results in a convex entropy minimization problem:

$$\begin{aligned} \underset{P \in \mathbb{S}^n}{\text{minimize}} \quad & \sum_i P_i \log\left(\frac{P_i}{Q_i^0}\right) \\ \text{subject to} \quad & P_i \leq \frac{\widehat{P}_i}{1-\alpha} \qquad i = 1, \ldots, n \end{aligned} \tag{5}$$

where $\alpha \in [0, 1]$ represents the fraction of samples removed.

More generally, $\mathcal{Q}$ is a set of distributions. The same continuous approximation results in a joint optimization over the model space $\mathcal{Q}$ and the space of empirical distributions, $\mathcal{P}(X, \alpha)$ defined in (2). Formally, let D_α^* be given as

$$D_\alpha^* = \min_{P \in \mathcal{P}(X,\alpha), Q \in \mathcal{Q}} \sum_i P_i \log\left(\frac{P_i}{Q_i}\right). \tag{6}$$

If $\mathcal{Q}$ is a convex set, the above optimization can be efficiently solved in many settings (see Sec. II-E).

To answer our original question and bound α^*, one can conduct a line search over $\alpha \in [0, 1]$, repeatedly solving the above optimization, and checking the output value of D_α^* against Theorem 1. This is captured in the following proposition.

Proposition 1. Let

$$\alpha_{\mathrm{L}} = \max\left\{ \alpha : D_\alpha^* \geq \frac{1}{p(1-\alpha)} \log\left(\frac{1}{\epsilon}\right) + \frac{2n}{p(1-\alpha)} \log\left(p(1-\alpha)+1\right) \right\} \tag{7}$$

then $\alpha_{\mathrm{L}} \leq \alpha^*$.

Proof. The proof follows directly from Theorem 1. For any α such that the condition on D_α^* in (7) holds, by Theorem 1, any distribution in $\mathcal{P}(X, \alpha)$ is contaminated. We note that

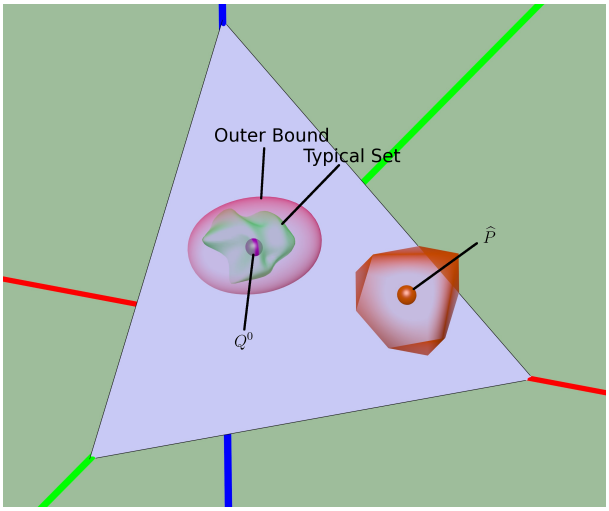

Fig. 1. Geometric interpretation of Proposition 1 and the optimization in (5) with $\mathcal{Q} = \{Q^0\}$. The width of the hypercube around $\widehat{P}$ is α. As α is increased, the hypercube eventually intersects the 'outer bound' set, which represents the set of distributions closest to Q^0 in KL divergence; the sets intersect when $\alpha = \alpha_{\mathrm{L}}$. Note that the 'outer' bound set also increases in size as α increases.

α_{L} always exists by monotone properties of D^*_α and the right hand side of the conditional in (7). See Appendix B, Theorem 6 for details. □

Fig. 1 shows a geometric interpretation of Proposition 1 and the optimization in (5). See the caption for details.

The lower bound obtained by solving the series of optimization problems converges to the separation distance, captured by the following theorem.

Theorem 2. *Let* $\mathcal{Q} = \{Q^0\}$. *Fix* $\widehat{P}(X)$. *Then*

$$\kappa(\widehat{P}||Q^0) - \alpha_{\mathrm{L}} = O\left(\sqrt{\frac{\log p}{p}}\right).$$

Proof. See Appendix B. □

Theorem 2 is stated for a fixed $\widehat{P}(X)$, although one would in general assume $\widehat{P}(X)$ to be an implicit function of p. The reason for fixing $\widehat{P}(X)$ is both generality and simplicity. The assumption decouples randomness from the convergence rate of the upper bound and the lower bound produced the optimization; without this assumption, the upper and lower bounds would be random variables, and necessitate a probabilistic statement. We also note that a precise limit statement can be readily extracted from the proof.

E. Discussion

In practice, it is often the case that the precise model distribution is not known; instead, it may be known that the model distribution comes from some family of distributions. This arises in anomaly detection when normal events are known to correspond to unknown proportions of samples from a finite set of distributions. This is the case of the mixture model *i.e.*, $\mathcal{Q}$ is the set of all distributions that can be represented as $Q = \sum \pi_j Q^j$ for any mixture proportions π_j. As the set of mixture distributions with unknown mixture components is a convex set, we can directly address this setting using the developments of Sec. II-D. Jointly optimizing over the mixture weights and the mixture distribution, the optimization takes the form

$$\underset{P\in\mathcal{P}^n,\ \pi\in\mathbb{S}^k}{\text{minimize}} \quad \sum_i P_i \log\left(\frac{P_i}{\sum_{j=1}^k \pi_j Q_i^j}\right). \tag{8}$$

We note that the above optimization can be solved at scale in real time for many applications; see discussions of numerical experiments below for details.

For many applications, model distributions are generated using a *finite* amount of data from known good sources (*i.e.*, sources that are known to have no contamination). Let $\widehat{Q}$ be an empirical distribution generated from p' samples of an $i.i.d.$ population, and consider the set

$$\mathcal{Q}' = \left\{Q : \widehat{Q} \text{ is typical for } (\{Q\}, \epsilon)\right\}.$$

Here, $\mathcal{Q}$ is the set of all distributions that have $\widehat{Q}$ as a typical empirical distribution. As before, determining membership in $\mathcal{Q}$ is intractable for large p' and more than two categories. Let

$$\bar{\mathcal{Q}} = \left\{Q : D(\widehat{Q}||Q) \le \frac{1}{p'}\log\left(\frac{1}{\epsilon}\right) + \frac{2n}{p'}\log(p'+1)\right\}.$$

$\bar{\mathcal{Q}}$ satisfies two important properties. First, $\mathcal{Q} \subseteq \bar{\mathcal{Q}}$ by Theorem 1 and second, $\bar{\mathcal{Q}}$ is a convex set.

Solving the optimization in (6) with $\mathcal{Q} = \bar{\mathcal{Q}}$ provides a powerful result which we state in the following proposition.

Proposition 2. Consider two empirical distributions $\widehat{P}$ and $\widehat{Q}$. Let $\mathcal{Q} = \bar{\mathcal{Q}}$, defined above, and let D^*_0 be the solution to the optimization in (6) with $\alpha = 0$. If

$$D^*_0 \ge \frac{1}{p}\log\left(\frac{1}{\epsilon}\right) + \frac{2n}{p}\log(p+1),$$

there is no Q that simultaneously satisfies *1)* $\widehat{Q}$ is typical with respect to Q and *2)* $\widehat{P}$ is typical with respect to Q.

Satisfying proposition 2 implies that observing a $\widehat{Q}$ and a $\widehat{P}$ generated by the same underlying distribution *by chance* can occur at most a fraction ϵ of the time; in this sense, $\widehat{P}$ must be contaminated. With a single parameter search over $\alpha \in [0,1]$, the lower bound applies: $\alpha^* \ge \alpha_{\mathrm{L}}$. We note that the formulation does not require the empirical model and the distribution under test to have joint support.

Numerical experiments were conducted to highlight the utility of Proposition 1; results are shown in Fig. 2. In contrast to the deterministic experiments in Fig. 2, experiments with random samples from various model and test distributions as input were run, showing similar convergence behavior. An experiment with with $\mathcal{Q}$ being a set of 10 mixture distributions with $n = 50$ was also conducted. The line search over α was completed using a bisecting search to an accuracy of 2^{-28} (the optimization was solved 27 times for each experiment). Averaged over 50 trials, the total time to compute α_{L} was 0.4 seconds. Experiments were implemented using CVXOPT [3] and results visualized with matplotlib [4].

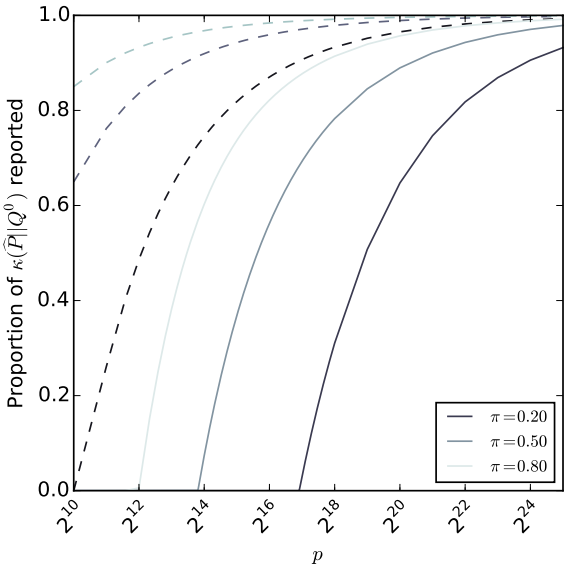

Fig. 2. Numerical example. $n = 11$, $\epsilon = 0.05$, $\mathcal{Q} = \{Q^0\}$, with Q^0 a uniform distribution over 11 categories. Solid lines show α_{L} divided by $\kappa(\widehat{P}||Q^0)$ for mixture distributions $\widehat{P}_{\text{dip}} = (1-\pi)\ Q^0 + \pi\mathcal{U}_{10}$, where $\mathcal{U}_{10}$ is a uniform distribution over 10 of the 11 categories. Dashed lines show α_L divided by $\kappa(\widehat{P}||Q^0)$ for $\widehat{P}_{\text{spike}} = (1-\pi)Q^0 + \pi\delta$, where δ is a point mass.

III. Related Work

Related work can be broadly classified into traditional work in goodness of fit (GoF) testing, and more recent work in anomaly detection. GoF testing has an extensive literature. When the data are binary valued, and the model distribution Bernoulli, quantifying contamination using GoF tests can be addressed by evaluating binomial probabilities (a technique known as Fisher's Exact method [5]). When the data take on more than two values, exact solutions for the level of contamination become intractable.

A customary approach to GoF testing for categorical data is Pearson's χ^2 test [6]. This approach to GoF testing can be quite powerful, but suffers from limitations. χ^2 tests are approximations, and are known to be invalid under certain conditions. In particular, the test is invalid when $p_i = 0$ for one or more categories. Nonetheless, employing the χ^2 test, one can deduce another optimization (much as we do in Sec. II-D) to answer the aforementioned question; we note the resulting optimization is a separable quadratic program with linear equality constraints which has an analytic solution [7], and would be an interesting starting point for future work. Since Pearson's χ^2 test hinges on a normal approximation, this approach would not result in strict contamination bounds. More specific to the contamination estimation problem presented here, recent work includes decontamination with multiclass label noise [8], [9], which focuses on recovering proportions of a set of mixture distributions present in dataset.

There is an extensive literature on the related topics of anomaly detection and outlier detection including work employing entropy based techniques, in particular [10] and [11]; we note the formulations here are distinct in that the level of contamination is not estimated. Lastly, we briefly discuss related work in anomaly detection the areas of computer networks, systems and security as this is the motivation for our developments. Early work on identifying anomalous or unexpected behaviors such faults (*e.g.,* due to outages or failures) or spikes (*e.g.,* associated with DoS attacks or flash crowds) in computer network traffic was based on the application of graph models, time series and multi-resolution methods *e.g.,* [12]–[15], and Principle Components Analysis (PCA) [16]–[18]. There are significant difficulties in tuning these methods to provide low false alarm rates in practice [19], necessitating methods based on statistical significance, as presented here.

References

[1] R. B. D'Agostino, *Goodness-of-fit-techniques*. CRC press, 1986.

[2] D. Aldous and P. Diaconis, "Strong uniform times and finite random walks," *Advances in Applied Mathematics*, vol. 8, no. 1, 1987.

[3] J. Dahl and L. Vandenberghe, "CVXOPT: Python software for convex optimization," 2011.

[4] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[5] C. R. Mehta, N. R. Patel, and A. A. Tsiatis, "Exact significance testing to establish treatment equivalence with ordered categorical data," *Biometrics*, pp. 819–825, 1984.

[6] A. Agresti, *Categorical data analysis*. John Wiley & Sons, 2014.

[7] L. BAY, J. Grau, M. Ruiz, and P. SU, "An analytic solution for some separable convex quadratic programming problems with equality and inequality constraints," *Journal of Mathematical Inequalities*, vol. 4, 2010.

[8] G. Blanchard and C. Scott, "Decontamination of mutually contaminated models," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014, pp. 1–9.

[9] C. Scott, G. Blanchard, and G. Handy, "Classification with asymmetric label noise: Consistency and maximal denoising." in *COLT*, ser. JMLR Proceedings, vol. 30, 2013.

[10] A. O. Hero, "Geometric entropy minimization (GEM) for anomaly detection and localization," in *Advances in Neural Information Processing Systems*, 2006, pp. 585–592.

[11] Y. Gu, A. McCallum, and D. Towsley, "Detecting anomalies in network traffic using maximum entropy estimation," in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*. USENIX Association, 2005, pp. 32–32.

[12] F. Feather, D. Siewiorek, and R. Maxion, "Fault Detection in an Ethernet Network Using Anomaly Signature Matching," in *Proceedings of ACM SIGCOMM*, San Francisco, CA, September 2000.

[13] I. Katzela and M. Schwartz, "Schemes for Fault Identification in Communications Networks," *IEEE/ACM Transactions on Networking*, vol. 3(6), pp. 753–764, December 1995.

[14] J. Brutlag, "Aberrant Behavior Detection in Time Series for Network Monitoring," in *Proceedings of the USENIX Fourteenth System Administration Conference LISA XIV*, New Orleans, LA, December 2000.

[15] P. Barford, J. Kline, D. Plonka, and A. Ron, "A Signal Analysis of Network Traffic Anomalies," in *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, Marseilles, France, November 2002.

[16] A. Lakina, M. Crovella, and C. Diot, "Diagnosing Network-wide Traffic Anomalies," in *Proceedings of ACM SIGCOMM*, Portland, OR, 2004.

[17] ——, "Characterization of Network-wide Anomalies in Traffic Flows," in *Proceedings of ACM Internet Measurement Conference*, Taormina, Italy, October 2004.

[18] ——, "Mining Anomalies Using Traffic Feature Distributions," in *Proceedings of ACM SIGCOMM*, Philadelphia, PA, August 2005.

[19] H. Ringberg, A. Soule, J. Rexford, and C. Diot, "Sensitivity of PCA for Traffic Anomaly Detection," in *Proceedings of ACM SIGMETRICS*, San Diego, CA, June 2007.

[20] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[21] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2009.

APPENDIX A

Proof of Theorem 1 requires two ingredients, both relying on results from large deviations theory. The first ingredient is Sanov's Theorem, which we state below.

Theorem 3. *(Sanov's Theorem) [20] (Theorem 11.4.1). Let $\mathcal{S}$ be a set of empirical distributions (with p samples over n categories). Then*

$$\mathbb{P}_Q(\mathcal{S}) \leq (p+1)^n \exp\left(-p \min_{\widehat{P}\in\mathcal{S}} D(\widehat{P}||Q)\right). \tag{9}$$

The second ingredient is also readily derived from results in large deviations theory.

Theorem 4. *Let $\mathcal{S}$ be a set of empirical distributions such that $\mathbb{P}_Q(\widehat{P}^\ell) \geq \mathbb{P}_Q(\widehat{P})$ for all $\widehat{P} \in \mathcal{S}$. Then,*

$$\min_{\widehat{P}\in\mathcal{S}} D(\widehat{P}||Q) \geq D(\widehat{P}^\ell||Q) - \frac{n}{p}\log(p+1).$$

Proof. The following inequalities hold [20] (Theorem 11.1.4):

$$\frac{1}{(p+1)^n}\exp\left(-pD(\widehat{P}||Q)\right) \leq \mathbb{P}_Q(\widehat{P}) \leq \exp\left(-pD(\widehat{P}||Q)\right). \tag{10}$$

Thus, for any $\mathbb{P}_Q(\widehat{P}^m) \leq \mathbb{P}_Q(\widehat{P}^\ell)$,

$$\frac{1}{(p+1)^n}\exp\left(-pD(\widehat{P}^m||Q)\right) \leq \exp\left(-pD(\widehat{P}^\ell||Q)\right)$$

which implies the result, completing the proof of Theorem 4. □

Combining Theorems 3 and 4, we have

$$\mathbb{P}_Q(\{\widehat{P}^1, \widehat{P}^2, \ldots, \widehat{P}^\ell\}) \leq (p+1)^{2n}\exp\left(-pD(\widehat{P}^\ell||Q)\right)$$

provided $\mathbb{P}_Q(\widehat{P}^1) \leq \mathbb{P}_Q(\widehat{P}^2) \leq \cdots \leq \mathbb{P}_Q(\widehat{P}^\ell)$. This provides a simple way to confirm if a sample is *contaminated* at a particular significance level ϵ. In particular, assume $\widehat{P}(X) = \widehat{P}^\ell$. If

$$(p+1)^{2n}\exp\left(-pD(\widehat{P}(X)||Q)\right) \leq \epsilon$$

or equivalently

$$D(\widehat{P}(X)||Q) \geq \frac{1}{p}\log\left(\frac{1}{\epsilon}\right) + \frac{2n}{p}\log(p+1) \tag{11}$$

then X is not typical; X satisfies

$$\mathbb{P}_Q(\{\widehat{P}^1, \widehat{P}^2, \ldots, \widehat{P}^\ell\}) \leq \epsilon$$

and is contaminated with significance ϵ. If (11) holds for all $Q \in \mathcal{Q}$, in other words, if

$$\inf_{Q\in\mathcal{Q}} D(\widehat{P}(X)||Q) \geq \frac{1}{p}\log\left(\frac{1}{\epsilon}\right) + \frac{2n}{p}\log(p+1)$$

we conclude then X is not typical with respect to $(\mathcal{Q}, \epsilon)$, implying the result.

APPENDIX B

Proof of Theorem 2. The proof requires three main steps. The first step is to show that when α is sufficiently close to $\kappa(\widehat{P}||Q^0)$, the solution to (6) can be written in closed form. The second step is to show a number of properties regarding the asymptotic behavior of α_{L} as p grows; specifically, α_{L} is monotone increasing in p, and converges to the separation distance; these properties imply that for large p, the closed form solution is valid. Lastly, we can bound the difference between $\kappa(\widehat{P}||Q^0)$ and α_{L} using the closed form solution.

Step 1: For α close to the separation distance (equivalently, for large p, as we show next in Theorem 6), the optimization has a closed form. This is captured in the following Theorem. Note the theorem assumes there is a unique largest $\frac{\widehat{P}_\ell}{Q_\ell}$; in the degenerate case when this is not true, the theorem can be restated introducing at most a factor of n, which does not affect the final result.

Theorem 5. *Let* $\frac{\widehat{P}_i}{Q_i}$ *be ordered such that* $\frac{\widehat{P}_\ell}{Q_\ell} < \frac{\widehat{P}_k}{Q_k} \leq \cdots \leq \frac{\widehat{P}_n}{Q_n}$. *For* $\alpha \in [1 - \widehat{P}_\ell - \frac{\widehat{P}_k}{Q_k}(1 - Q_\ell),\ \kappa(\widehat{P}||Q)]$

$$P_i^* = \begin{cases} \frac{Q_i\left(1 - \frac{\widehat{P}_\ell}{1-\alpha}\right)}{1-Q_\ell} & i \neq \ell \\ \frac{\widehat{P}_\ell}{1-\alpha} & i = \ell \end{cases} \tag{12}$$

is the unique solution to (6).

Proof. The result can be shown by verifying the conditions KKT conditions with

$$\lambda_i^* = \begin{cases} 0 & i \neq \ell \\ \log\left(\frac{Q_\ell\left(1 - \frac{\widehat{P}_\ell}{1-\alpha}\right)}{(1-Q_\ell)\frac{\widehat{P}_\ell}{1-\alpha}}\right) & i = \ell \end{cases} \tag{13}$$

and

$$\nu^* = \log\left(\frac{1 - Q_\ell}{1 - \frac{\widehat{P}_\ell}{1-\alpha}}\right) - 1$$

where the Lagrangian [21] is given as

$$L(P, \lambda, \nu) = \sum_i P_i \log \frac{P_i}{Q_i} + \sum_i \lambda_i \left(P_i - \frac{\widehat{P}_i}{1-\alpha}\right) + \nu\left(\sum_i P_i - 1\right).$$

These primal and dual optimal points are derived using methods similar to [21] (p. 228, 248); in what follows, we simply verify the KKT conditions which suffice to complete the proof. First, we confirm that the solution is a stationary point:

$$\left.\frac{\partial L(P, \lambda, \nu)}{\partial P_i}\right|_{P^*, \lambda^*, \nu^*} = \left.\log \frac{P_i}{Q_i} + 1 + \lambda_i + \nu\right|_{P^*, \lambda^*, \nu^*} = 0$$

which holds for all i. The complementary slackness condition is readily verified:

$$\lambda_i^*\left(P_i^* - \frac{\widehat{P}_i}{1-\alpha}\right) = 0, \quad i = 1, \ldots, n.$$

It remains to show conditions under which the solution is primal and dual feasible. First, $\lambda_\ell^* \geq 0$ provided

$$\frac{Q_\ell\left(1 - \frac{\widehat{P}_\ell}{1-\alpha}\right)}{(1 - Q_\ell)\frac{\widehat{P}_\ell}{1-\alpha}} \geq 1.$$

After arranging terms, the above holds when $\alpha \leq 1 - \frac{\widehat{P}_\ell}{Q_\ell} = \kappa(\widehat{P}||Q)$. The primal equality constraint, $\sum_i P_i^* = 1$, is readily verified. Lastly, we check the primal inequality constraints. P_ℓ^* is trivially feasible. For $i \neq \ell$, we require

$$P_i^* = \frac{Q_i\left(1 - \frac{\widehat{P}_\ell}{1-\alpha}\right)}{1 - Q_\ell} \leq \frac{\widehat{P}_i}{1-\alpha}$$

which holds when

$$\alpha \geq 1 - \widehat{P}_\ell - \frac{\widehat{P}_i}{Q_i}(1 - Q_\ell).$$

Since $\frac{\widehat{P}_i}{Q_i} \geq \frac{\widehat{P}_k}{Q_k}$ for all $i \neq \ell$, the solution is feasible if

$$\alpha \geq 1 - \widehat{P}_\ell - \frac{\widehat{P}_k}{Q_k}(1 - Q_\ell).$$

We conclude that the KKT conditions are satisfied for the range α specified in the statement of the theorem. Since the objective is strictly convex the solution is unique, which completes the proof.

□

Step 2: We show that as p approaches infinity, α approaches the separation distance. More specifically, we have the following theorem.

Theorem 6.

$$\alpha_{\mathrm{L}} \leq \kappa(\widehat{P}||Q^0) \qquad \textit{and} \qquad \lim_{p\to\infty} \alpha_{\mathrm{L}} = \kappa(\widehat{P}||Q^0)$$

We begin the proof by examining the behavior of D^*_α and α_{L}. Note that D^*_α (the minimizer of (6)) is monotone non-increasing in α, as increasing α relaxes the constraints. For $\alpha = \kappa(\widehat{P}||Q^0)$, $D^*_\alpha = 0$ as the constraints allow $P_i = Q_i$ for all i (as KL divergence is minimized if and only if $P_i = Q_i$ for all i). Define

$$\gamma_{\mathrm{L}}(\alpha, p) = \frac{1}{p(1-\alpha)} \log\left(\frac{1}{\epsilon}\right) + \frac{2n}{p(1-\alpha)} \log\left(p(1-\alpha)+1\right)$$

for $\alpha \in [0,1]$, $p > 0$. We can write (7) as

$$\alpha_{\mathrm{L}} = \max\left\{\alpha : D^*_\alpha \geq \gamma_{\mathrm{L}}(\alpha, p)\right\}$$

For fixed p, $\gamma_{\mathrm{L}}(\alpha, p)$ is strictly increasing in α for $\alpha \in [0,1]$. This (and since D^*_α is monotone non-decreasing in α) implies existence and uniqueness of α_{L} for fixed p. Next, for fixed α, $\gamma_{\mathrm{L}}(\alpha, p)$ is strictly decreasing in p. Since D^*_α is not a function of p, we conclude that α_{L} is non-decreasing in p.

Lastly, to prove the limit statement, we require D^*_α be left continuous at $\alpha = \kappa(\widehat{P}||Q^0)$; for any $\epsilon > 0$, there exists some $\delta > 0$ such that $D^*_{\kappa(\widehat{P}||Q^0)-\delta} < \epsilon$. This follows as the objective is continuous in the optimization variables, and constraints are continuous in α; an arbitrarily small increase in the objective can be realized by sufficiently reducing α.

Step 3: Bound $\kappa(\widehat{P}||Q) - \alpha_{\mathrm{L}}$ using the closed form solution.

The value of KL divergence at P^* from (12) is

$$\begin{aligned}
D^*_\alpha &= D(P^*||Q) = \sum_{i\neq\ell} \frac{Q_i\left(1-\frac{\widehat{P}_\ell}{1-\alpha}\right)}{1-Q_\ell} \log\left(\frac{1-\frac{\widehat{P}_\ell}{1-\alpha}}{1-Q_\ell}\right) + \frac{\widehat{P}_\ell}{1-\alpha} \log\frac{\frac{\widehat{P}_\ell}{1-\alpha}}{Q_\ell} \\
&= \left(1-\frac{\widehat{P}_\ell}{1-\alpha}\right) \log\left(\frac{1-\frac{\widehat{P}_\ell}{1-\alpha}}{1-Q_\ell}\right) + \frac{\widehat{P}_\ell}{1-\alpha} \log\frac{\frac{\widehat{P}_\ell}{1-\alpha}}{Q_\ell} \\
&\leq \frac{\left(Q_\ell - \frac{\widehat{P}_\ell}{1-\alpha}\right)^2}{Q_\ell(1-Q_\ell)} \\
&= \frac{Q_\ell\left(\kappa(\widehat{P}||Q)-\alpha\right)^2}{(1-Q_\ell)(1-\alpha)^2} \qquad (14)
\end{aligned}$$

where the inequality follows since $\log(x) \leq x-1$. We are ready to bound the difference between α_{L} and the separation distance. Recall the definition of α_{L}; α_{L} must satisfy

$$D^*_{\alpha_{\mathrm{L}}} \leq \frac{1}{p(1-\alpha_{\mathrm{L}})} \log\left(\frac{1}{\epsilon}\right) + \frac{n}{p(1-\alpha_{\mathrm{L}})} \log\left(p(1-\alpha_{\mathrm{L}})+1\right)$$

and by (14)

$$\frac{Q_\ell\left(\kappa(\widehat{P}||Q)-\alpha_{\mathrm{L}}\right)^2}{(1-\alpha_{\mathrm{L}})(1-Q_\ell)} \leq \frac{1}{p}\log\left(\frac{1}{\epsilon}\right) + \frac{n}{p}\log\left(p(1-\alpha_{\mathrm{L}})+1\right)$$

which implies the result

$$\kappa(\widehat{P}||Q) - \alpha_{\mathrm{L}} \leq \sqrt{\frac{1}{p}\log\left(\frac{1}{\epsilon}\right) + \frac{n}{p}\log\left(p+1\right)} = O\left(\sqrt{\frac{\log p}{p}}\right).$$