Active Diagnosis under Persistent Noise with Unknown Noise Distribution: A Rank-Based Approach

Gowtham Bellala[*], **Suresh K. Bhavnani**[†], **Clayton Scott**[*]
[*]University of Michigan, Ann Arbor, MI, [†]University of Texas Medical Branch, Galveston, TX
gowtham@umich.edu, skbhavnani@gmail.com, clayscot@umich.edu

Abstract

We consider a problem of active diagnosis, where the goal is to efficiently identify an unknown object by sequentially selecting, and observing, the responses to binary valued queries. We assume that query observations are noisy, and further that the noise is persistent, meaning that repeating a query does not change the response. Previous work in this area either assumed the knowledge of the query noise distribution, or that the noise level is sufficiently low so that the unknown object can be identified with high accuracy. We make no such assumptions, and introduce an algorithm that returns a ranked list of objects, such that the expected rank of the true object is optimized. Furthermore, our algorithm does not require knowledge of the query noise distribution.

1 Introduction

We study an active diagnosis problem where the goal is to identify an unknown object while minimizing the number of binary questions posed about that object. This problem arises in various places such as pool-based active learning (Dasgupta, 2004; Nowak, 2008; Golovin and Krause, 2010), disease diagnosis (Loveland, 1985; Yu et al., 2009), fault diagnosis in computer networks (Rish et al., 2005), toxic chemical identification (Bhavnani et al., 2007), image processing (Korostelev and Kim, 2000), computer vision (Swain and Stricker, 1993; Geman and Jedynak, 1996), job scheduling (Kosaraju et al., 1999), and the adaptive traveling salesperson problem (Gupta et al., 2010). These problems can be characterized in terms of a set $\Theta = \{\theta_1, \cdots, \theta_M\}$ of M different objects and a set $Q = \{q_1, \cdots, q_N\}$ of N distinct subsets of Θ known as queries. An unknown object $\theta \in \Theta$ is generated with a certain *a priori* probability distribution $\Pi = (\pi_1, \cdots, \pi_M)$, i.e., $\pi_i = \Pr(\theta = \theta_i)$, and the goal is to identify θ through as few queries as possible, where a query $q \in Q$ returns a value of 1 if $\theta \in q$, and 0 otherwise. In many applications, the responses to queries are corrupted by noise. For example, in active learning, the objects are classifiers, queries are labels to fixed test points, and the noise is due to a faulty oracle. Similarly, in fault diagnosis, objects may correspond to components, queries to alarms, and the noise is either due to unreliable connections or faulty devices.

The problem of active diagnosis/active learning in the presence of query noise has been studied by Kääriäinen (2006) and Nowak (2008, 2009), where the noise is assumed to be independent, in that posing the same query twice may yield different responses. This assumption suggests repeated selection of a query as a possible strategy to overcome query noise. The algorithms presented in (Kääriäinen, 2006; Nowak, 2008, 2009) are based on this principle. However, in certain applications, resampling or repeating a query does not change the query response, thereby confining an active diagnosis algorithm to non-repeatable queries.

For example, in the emergency response problem of toxic chemical identification (Bhavnani et al., 2007), a first responder is faced with the task of rapidly identifying the toxic chemical by posing symptom-based queries to a victim. The responses to these symptom queries are often in error due to reasons such as mis-identification of a symptom by a victim or a delayed onset of a symptom, in which case the victim's response is unlikely to change upon repeated queries. Similarly, in a fault diagnosis problem, the response to alarms/probes could be in error due to faulty alarms, in which case these responses would not change on repeated interrogations.

This more stringent noise model where queries cannot be resampled is referred to as persistent noise (Rényi, 1961; Hanneke, 2007). It has been studied earlier in the situation where the number of persistent errors is restricted such that unique identification of the unknown object θ is guaranteed (Bellala et al., 2009; Golovin et al., 2010). In par-

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

ticular, if we associate each object θ_i with a length N bit string whose j^{th} element is 1 iff $\theta_i \in q_j$, then, the number of query errors is restricted to be less than half of the minimum Hamming distance between any two object bit strings. This is often not reasonable as the minimum Hamming distance could be very small, such as in WISER[1] (a toxic chemical database) where it is equal to 1.

In this paper, we consider the problem of active diagnosis under persistent noise with no restriction on the number of persistent errors. We assume the object set Θ and the query set Q are finite, and that only one object from Θ is "present". Unlike the previous two noise models where the unknown object θ can be identified with certainty after sufficiently many queries, in our model it may not be possible to identify θ even after all queries are made.

In this setting, Rish et al. (2005) proposed the use of mutual information or the conditional entropy as a criterion for selecting queries, where queries are chosen sequentially to minimize the uncertainty in θ (or maximize information gain) given the observed responses to the past queries. After observing responses to a set of queries, the unknown object is then estimated to be the object with the maximum *a posteriori* probability, θ_{MAP}.

However, there are two limitations with this approach. First, in situations with moderate to high noise, or where the Hamming distance between object bit strings is low, the object with the maximum *a posteriori* probability will be equal to the true object θ with low probability. Even in the case where θ_{MAP} does converge to the true object θ, it may require a large number of queries to be inputted. Second, this algorithm assumes knowledge of the full data model; in particular, the probability of query errors, which is required to compute the information gain in the query selection stage. However, this information is often not known.

To address these issues, we propose a novel rank-based approach where we output a ranked list of objects rather than θ_{MAP}, where the ranking is based on the posterior probabilities. The rank-based approach is motivated by the fact that in many applications there is a domain expert who makes the final decision on the possible identity of the unknown object θ. Such a ranking can be useful to a domain expert who will use domain expertise and other sources of information to further determine θ. Thus, we propose a greedy algorithm to minimize the expected rank of the unknown object θ. Unlike the entropy-based algorithm, the proposed greedy algorithm does not require the knowledge of the underlying query noise.

1.1 Additional Related Work

In the noise-free case, this problem has been referred to as binary testing or object/entity identification (Garey, 1972; Loveland, 1985). The goal of object identification is to construct an optimal binary decision tree where each internal node in the tree is associated with a query from Q, and each leaf corresponds to an object from Θ, where optimality is often with respect to the average depth of the leaf nodes. This problem of finding an optimal decision tree is known to be NP-complete (Hyafil and Rivest, 1976). However, there exists an efficient, greedy algorithm known as the *splitting algorithm* or *generalized binary search* (GBS) that achieves a logarithmic approximation to the optimal solution (Dasgupta, 2004; Nowak, 2008).

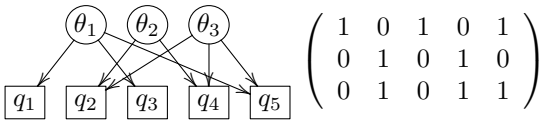

Figure 1: A bipartite graph along with a binary matrix representation of a dataset with 3 objects and 5 queries where $q_1 = q_3 = \{\theta_1\}$, $q_2 = q_4 = \{\theta_2, \theta_3\}$ and $q_5 = \{\theta_1, \theta_3\}$.

The problem of pool-based active learning under persistent noise has been studied by Balcan et al. (2006) and Hanneke (2007) in the PAC (Probably Approximately Correct) model. The query set is assumed to be large enough (possibly infinite) such that it is possible to get arbitrarily close to the optimal classifier, for any given noise level.

2 Data model

We derive our rank-based algorithm under a very flexible data model. We present this general model because it encompasses two important cases as discussed in Section 5, which are the main focus of this paper. In each of those cases, we show that the proposed algorithm can be implemented without any knowledge on the model parameters.

Given an object set $\Theta = \{\theta_1, \cdots, \theta_M\}$ of M different objects and a query set $Q = \{q_1, \cdots, q_N\}$ of N distinct subsets of Θ, the relation between the objects and queries can be represented either by a bipartite network or by an $M \times N$ binary matrix $\mathbf{B} = [b_{ij}]$, where $b_{ij} = 1$ if $\theta_i \in q_j$, and 0 otherwise. Each row of $\mathbf{B}$ is an object bit string, defined previously. Figure 1 demonstrates a bipartite graph representation for a toy dataset along with the binary matrix associated with it.

We associate each object $\theta_i \in \Theta$ with a binary random variable X_i, where $X_i = 1$ when $\theta = \theta_i$, and 0 otherwise. Then, $\mathbf{X} = (X_1, \cdots, X_M)$ is a binary random vector denoting the states of all the objects in Θ, where $\mathbf{X} \in \{\mathbb{I}_1, \cdots, \mathbb{I}_M\}$, $\mathbb{I}_i$ being a binary vector whose ith element is 1 and remaining elements are 0, and $\Pr(\mathbf{X} = \mathbb{I}_i) = \pi_i$.

Similarly, let Z_j be a binary random variable denoting the observed response to query q_j. Then, $\mathbf{Z} = (Z_1, \cdots, Z_N)$ is a binary random vector denoting the observed responses to

[1] http://wiser.nlm.nih.gov/

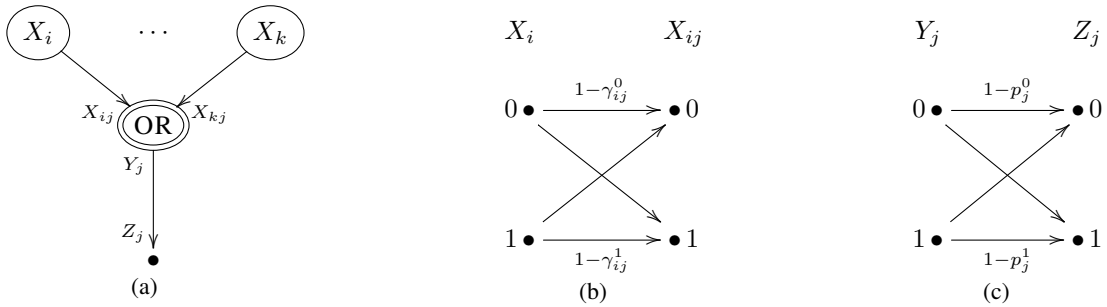

Figure 2: (a) Noise model (b) Unreliable connections (c) Noisy query responses

all queries in Q, where $\mathbf{Z} \in \{0,1\}^N$. We wish to define the joint distribution of $(\mathbf{X}, \mathbf{Z})$, for which it remains to specify the conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$. Towards this end, we present the model describing the relation between an observed response Z_j to the object state vector $\mathbf{X}$.

For any query q_j, let $\mathbf{pa}_j := \{i : b_{ij} = 1\}$ denote the objects that are connected to it. In the ideal case when there is no noise, the response Z_j of query q_j is given by the OR operation of the binary states of the objects in $\mathbf{pa}_j$. More generically, the response of a query q_j can be modeled as shown in Figure 2(a). This noise model has been referred to as the Y model (Le and Hadjicostis, 2007). Here, for any $i \in \mathbf{pa}_j$, X_{ij} denotes the binary state of object θ_i as perceived at query q_j. This corresponds to randomness due to unreliable connections in a fault-diagnosis problem. Furthermore, Y_j which denotes the outcome of the OR operation on $\{X_{ij}\}_{i \in \mathbf{pa}_j}$, is observed as Z_j. The relationship between Y_j and Z_j corresponds to noise due to unreliable alarms in a fault-diagnosis problem or a faulty oracle in an active learning setting. This noise model can be completely characterized by the parameters $0 \leq \gamma_{ij}^x, p_j^x \leq 1$, $x = 0, 1$, as shown in Figures 2(b) and 2(c). For any query $q_j \in Q$, and $x = 0, 1$,

$$\gamma_{ij}^x = \Pr(X_{ij} = 1 - x | X_i = x),\ \forall i \in \mathbf{pa}_j$$
$$\text{and } p_j^x = \Pr(Z_j = 1 - x | Y_j = x).$$

Finally, let $Q_{\mathcal{A}}$ denote the subset of queries indexed by $\mathcal{A} \subseteq \{1, \cdots, N\}$, and $\mathbf{Z}_{\mathcal{A}}$ the random variables associated with those queries, e.g, if $\mathcal{A} = \{1, 4, 7\}$, then $Q_{\mathcal{A}} = \{q_1, q_4, q_7\}$ and $\mathbf{Z}_{\mathcal{A}} = (Z_1, Z_4, Z_7)$. Then, we make the standard assumption that the observed responses to queries are conditionally independent given the states of all the objects, i.e.,

$$\Pr(\mathbf{Z}_{\mathcal{A}} = \mathbf{z}_{\mathcal{A}} | X_i = 1) = \prod_{j \in \mathcal{A}} \Pr(Z_j = z_j | X_i = 1),$$

where by $X_i = 1$, we implicitly mean $\mathbf{X} = \mathbb{I}_i$. This assumption holds reasonably well in many practical applications as noise is usually generated independently. For example, in the problem of fault diagnosis, it can be reasonable to assume that all connections and alarms fail independently.

3 Active Diagnosis under Persistent noise

We will now formally state the problem of active diagnosis. As mentioned earlier, unique identification of θ is no longer guaranteed. Hence, the goal of active diagnosis under persistent noise is to maximize some function $f(Q_{\mathcal{A}}, \mathbf{z}_{\mathcal{A}})$, which denotes the quality of the estimate of θ, subject to a constraint on the number of queries made, i.e.,

$$\max_{\mathcal{A} \subseteq \{1, \cdots, N\}} f(Q_{\mathcal{A}}, \mathbf{z}_{\mathcal{A}})$$
$$\text{s.t. } |\mathcal{A}| \leq k.$$

Finding an optimal solution to this problem is NP-complete. Instead, the queries can be chosen sequentially by greedily maximizing the quality function, given the observed responses to the past queries, i.e.,

$$q^* := \operatorname*{argmax}_{q \in Q \setminus Q_{\mathcal{A}}} \mathbb{E}_Z[f(Q_{\mathcal{A}} \cup \{q\}, [\mathbf{z}_{\mathcal{A}}, Z]) - f(Q_{\mathcal{A}}, \mathbf{z}_{\mathcal{A}}) | \mathbf{Z}_{\mathcal{A}} = \mathbf{z}_{\mathcal{A}}] \quad (1)$$

where Z is the random variable associated with query q.

In the case of entropy-based active diagnosis, this quality function is given by $f(Q_{\mathcal{A}}, \mathbf{z}_{\mathcal{A}}) = H(\theta) - H(\theta | \mathbf{z}_{\mathcal{A}})$, which is the reduction in conditional entropy, or the information gain. Given the observed responses $\mathbf{z}_{\mathcal{A}}$ to previously selected queries $Q_{\mathcal{A}}$, the next query is chosen to be

$$q^* = \operatorname*{argmin}_{q \in Q \setminus Q_{\mathcal{A}}} \sum_{z=0,1} \Pr(Z = z | \mathbf{z}_{\mathcal{A}}) H(\mathbf{X} | \mathbf{z}_{\mathcal{A}}, z)$$

where the conditional entropy $H(\mathbf{X} | \mathbf{z}_{\mathcal{A}}, z)$ is given by

$$-\sum_{i=1}^{M} \Pr(X_i = 1 | \mathbf{z}_{\mathcal{A}}, z) \log_2(\Pr(X_i = 1 | \mathbf{z}_{\mathcal{A}}, z)).$$

The computation of these posterior probabilities requires the knowledge of the complete noise distribution or the parameters in the noise model. In the next section, we propose a rank-based greedy algorithm that depends instead on the likelihoods and the prior probability distribution. We then exploit this fact in Section 5 to develop algorithms that do not require knowledge of the query noise parameters.

4 Rank-Based Active Query Selection

Given the observed responses $\mathbf{z}_\mathcal{A}$ to a set of queries $Q_\mathcal{A}$, we define the rank of an object θ_i to be

$$R(\theta_i|\mathbf{z}_\mathcal{A}) = \sum_{j=1}^{M} \mathbf{I}\Big\{\Pr(X_j = 1|\mathbf{z}_\mathcal{A}) \geq \Pr(X_i = 1|\mathbf{z}_\mathcal{A})\Big\}$$
$$= \sum_{j=1}^{M} \mathbf{I}\Big\{\pi_j \Pr(\mathbf{z}_\mathcal{A}|X_j = 1) \geq \pi_i \Pr(\mathbf{z}_\mathcal{A}|X_i = 1)\Big\},$$

where $\mathbf{I}\{E\}$ is an indicator function which takes the value 1 when the event E is true, and 0 otherwise. Note that $R(\theta_i|\mathbf{z}_\mathcal{A})$ takes a small value when θ_i has a high posterior probability and a large value when the posterior probability is small. In addition, when multiple objects have the same posterior probabilities, each object is assigned the worst case ranking, as shown in Figure 3.

Given the observed responses $\mathbf{z}_\mathcal{A}$ to a set of queries $Q_\mathcal{A}$, we define the objective function $f(Q_\mathcal{A}, \mathbf{z}_\mathcal{A})$ to be the expected rank of the unknown object θ, i.e.,

$$\mathbb{E}_\theta[R(\theta|\mathbf{z}_\mathcal{A})] = \sum_{i=1}^{M} \Pr(X_i = 1|\mathbf{z}_\mathcal{A}) R(\theta_i|\mathbf{z}_\mathcal{A}), \quad (2)$$

and the goal is to minimize this expected rank. Substituting this objective function in (1), we get the criterion for choosing the next query to be

$$q^* = \operatorname*{argmin}_{q \in Q \setminus Q_\mathcal{A}} \sum_{z=0,1} \Pr(Z = z|\mathbf{z}_\mathcal{A}) \mathbb{E}_\theta[R(\theta|\mathbf{z}_\mathcal{A}, z)]$$
$$= \operatorname*{argmin}_{q \in Q \setminus Q_\mathcal{A}} \sum_{z=0,1} \sum_{i=1}^{M} \frac{\pi_i \Pr(\mathbf{z}_\mathcal{A}, z|X_i = 1)}{\Pr(\mathbf{z}_\mathcal{A})} R(\theta_i|\mathbf{z}_\mathcal{A}, z)$$
$$= \operatorname*{argmin}_{q \in Q \setminus Q_\mathcal{A}} \sum_{z=0,1} \sum_{i=1}^{M} \pi_i \Pr(\mathbf{z}_\mathcal{A}, z|X_i = 1) R(\theta_i|\mathbf{z}_\mathcal{A}, z) \quad (3)$$

where (3) follows as $\Pr(\mathbf{z}_\mathcal{A})$ does not depend on query q. Given the knowledge of γ_{ij}^x, p_j^x, and π_i, the greedy algorithm can now be implemented. In the noise-free case with uniform prior on the objects, this rank-based greedy algorithm reduces to GBS.

5 Noise Independent Active Query Selection

We now consider two special cases of the noise model discussed in Section 2 that appear in many applications, and present a noise independent estimate of the objective in (3). More specifically, we provide a good upper bound on the likelihood function, which can then be used to accurately predict the ranks of the objects. We also show that in some cases it is possible to estimate the true ranks exactly with limited knowledge on the query noise. We use the result in the following lemma to derive the upper bound on the likelihood function.

	θ_1	θ_2	θ_3	θ_4	θ_5
$\Pr(X_i = 1\|\mathbf{z}_\mathcal{A})$	0.2	0.2	0.3	0.2	0.1
$R(\theta_i\|\mathbf{z}_\mathcal{A})$	4	4	1	4	5

Figure 3: Demonstration of worst case ranking

Lemma 1. *Let h, k be integers with $0 \leq h \leq k$ and $k \geq 1$. Then, for any $0 < p < 1$,*

$$p^h(1-p)^{k-h} \leq \varepsilon_h^h (1-\varepsilon_h)^{k-h} \quad (4)$$

where $\varepsilon_h = \frac{h}{k}$. If it is known that $p \leq p_2 < 1$, then (4) *holds with $\varepsilon_h = \min\{p_2, \frac{h}{k}\}$. If it is known that $p \geq p_1 > 0$, then* (4) *holds with $\varepsilon_h = \max\{p_1, \frac{h}{k}\}$. If it is known that $0 < p_1 \leq p \leq p_2 < 1$, then* (4) *holds with $\varepsilon_h = \min\{p_2, \max\{p_1, \frac{h}{k}\}\}$.*

Proof. For any given k and h, let $g(p) := \log[p^h(1-p)^{k-h}]$. It can be easily verified that $g'(p) = 0$ when $p = \frac{h}{k}$ and $g''(p)|_{p=\frac{h}{k}} < 0$ which implies that $g(p) \leq g(\frac{h}{k})$, $\forall\, p$, from which the inequality in (4) follows.

In addition, when $p \leq p_2$, we need to show that the bound can be improved to

$$p^h(1-p)^{k-h} \leq \begin{cases} p_2^h(1-p_2)^{k-h} & \text{if } p_2 \leq \frac{h}{k}, \\ \left(\frac{h}{k}\right)^h \left(1-\frac{h}{k}\right)^{k-h} & \text{if } p_2 > \frac{h}{k}. \end{cases}$$

Note that the second part of this result, where $p_2 > h/k$ follows from the above result. Hence, it remains to show that $\forall\, p_2 \leq \frac{h}{k}$, $p^h(1-p)^{k-h} \leq p_2^h(1-p_2)^{k-h}$, which is equivalent to showing that $\forall h \geq kp_2$, $g(p_2) - g(p) \geq 0$.

$$g(p_2) - g(p) = h \log \frac{p_2(1-p)}{p(1-p_2)} + k \log \frac{1-p_2}{1-p}$$
$$\geq kp_2 \log \frac{p_2(1-p)}{p(1-p_2)} + k \log \frac{1-p_2}{1-p}$$
$$= k\left[p_2 \log \frac{p_2}{p} + (1-p_2) \log \frac{1-p_2}{1-p}\right] \geq 0$$

where the first inequality follows from $h \geq kp_2$ (the first log is ≥ 0 since $p \leq p_2$) and the last inequality follows from the non-negativity of Kullback-Leibler divergence. The other two cases can be proved in a similar manner. □

5.1 Constant Noise Level

We begin with the following special case of the noise model, where $\gamma_{ij}^x = 0$, $\forall i, j, x$ and $0 < p_j^x = p < 1$, $\forall j, x$. This noise model has been used in the context of pool-based active learning with a faulty oracle (Hanneke, 2007; Nowak, 2009), experimental design (Rényi, 1961), computer vision, and image processing (Korostelev and Kim, 2000), where the responses to some queries are assumed to be randomly flipped. In this setting,

$$\Pr(Z_j = z_j|X_i = 1) = p^{|b_{ij} - z_j|}(1-p)^{1-|b_{ij}-z_j|}.$$

More generally,

$$\Pr(\mathbf{Z}_{\mathcal{A}} = \mathbf{z}_{\mathcal{A}}|X_i = 1) = p^{\delta_{i,\mathcal{A}}}(1-p)^{|\mathcal{A}|-\delta_{i,\mathcal{A}}},$$

where $\delta_{i,\mathcal{A}} = \sum_{j\in\mathcal{A}} |b_{ij} - z_j|$, is the local Hamming distance between the true responses of θ_i to queries in $Q_{\mathcal{A}}$, and the observed responses $\mathbf{z}_{\mathcal{A}}$. Using the result in Lemma 1, the above likelihood function can be upper bounded by

$$\overline{\Pr}(\mathbf{z}_{\mathcal{A}}|X_i = 1) := \left(\frac{\delta_{i,\mathcal{A}}}{|\mathcal{A}|}\right)^{\delta_{i,\mathcal{A}}} \left(1 - \frac{\delta_{i,\mathcal{A}}}{|\mathcal{A}|}\right)^{|\mathcal{A}|-\delta_{i,\mathcal{A}}}.$$

Note that the lemma also states that given any additional information on the noise parameter p, this bound can be further improved. Let $\overline{R}(\theta_i|\mathbf{z}_{\mathcal{A}})$ denote the estimated rank of object θ_i based on these upper bounds:

$$\overline{R}(\theta_i|\mathbf{z}_{\mathcal{A}}) := \sum_{j=1}^{M} \mathbf{I}\Big\{\pi_j \overline{\Pr}(\mathbf{z}_{\mathcal{A}}|X_j = 1) \geq \pi_i \overline{\Pr}(\mathbf{z}_{\mathcal{A}}|X_i = 1)\Big\}, \quad (5)$$

Then, the query selection criterion in (3) can be replaced by the following noise-independent criterion

$$\operatorname*{argmin}_{q\in Q\setminus Q_{\mathcal{A}}} \sum_{z=0,1}\sum_{i=1}^{M} \pi_i \overline{\Pr}(\mathbf{z}_{\mathcal{A}}, z|X_i = 1)\overline{R}(\theta_i|\mathbf{z}_{\mathcal{A}}, z). \quad (6)$$

The result in Proposition 1 presents conditions under which the true rank can be estimated accurately. It states that, under uniform prior on the objects, it suffices to know whether $p < 0.5$ or $p > 0.5$, for the estimated ranks to be exactly equal to the true ranks.

More generally, for any given prior Π with $\rho := \min_i \pi_i / \max_i \pi_i$, it suffices to know whether $p < \frac{\rho}{1+\rho}$ or $p > \frac{1}{1+\rho}$, for the estimated ranks to be equal to the true ranks. However, even in the case where $\frac{\rho}{1+\rho} \leq p < 0.5$ or $0.5 < p \leq \frac{1}{1+\rho}$, it turns out that the estimated ranks will be equal to the true ranks for most of the objects, as demonstrated by the experiments in the Supplemental[2].

Proposition 1. *Let $\mathbf{z}_{\mathcal{A}}$ be the observed responses to a sequence of queries $Q_{\mathcal{A}} \subseteq Q$, under some unknown noise parameter p. Let $\rho := \min_i \pi_i / \max_i \pi_i$. Given a $\overline{p} \in (0, \frac{\rho}{1+\rho})$ such that $0 < p \leq \overline{p}$, or a $\underline{p} \in (\frac{1}{1+\rho}, 1)$ such that $1 > p \geq \underline{p}$, the estimated ranks $\overline{R}(\theta|\mathbf{z}_{\mathcal{A}})$ computed only with the knowledge of $\overline{p}$ or $\underline{p}$ are equal to the true ranks $R(\theta|\mathbf{z}_{\mathcal{A}})$, $\forall\, \theta \in \Theta$.*

Proof. Let $|\mathcal{A}| = k$. Consider the case where $\exists\ \overline{p} \in (0, \rho/(1+\rho))$ such that $0 < p \leq \overline{p}$ (The other case where $\exists\ \underline{p} \in (1/(1+\rho), 1)$ such that $1 > p \geq \underline{p}$ can be proved in a similar manner). Note from the definitions of $R(\theta|\mathbf{z}_{\mathcal{A}})$ and $\overline{R}(\theta|\mathbf{z}_{\mathcal{A}})$ that the result follows by showing the following relational equivalence between the true probabilities and the estimated probabilities: $\forall i, j$

$$\pi_i \Pr(\mathbf{z}_{\mathcal{A}}|X_i = 1) \geq \pi_j \Pr(\mathbf{z}_{\mathcal{A}}|X_j = 1) \iff \pi_i \overline{\Pr}(\mathbf{z}_{\mathcal{A}}|X_i = 1) \geq \pi_j \overline{\Pr}(\mathbf{z}_{\mathcal{A}}|X_j = 1), \quad (7)$$

where the true likelihood and the estimated likelihood of any object θ_i are given by $\Pr(\mathbf{z}_{\mathcal{A}}|X_i = 1) = p^{h_i}(1-p)^{k-h_i}$ and $\overline{\Pr}(\mathbf{z}_{\mathcal{A}}|X_i = 1) = \varepsilon_i^{h_i}(1-\varepsilon_i)^{k-h_i}$, $h_i = \delta_{i,\mathcal{A}}$ and $\varepsilon_i := \min\{h_i/k, \overline{p}\}$.

The above equivalence follows trivially for any pair of objects θ_i, θ_j whose $h_i = h_j$. To show that the equivalence holds even when $h_i \neq h_j$, we will show that, for any two objects θ_i, θ_j with priors π_i, π_j,

$$\pi_i \Pr(\mathbf{z}_{\mathcal{A}}|X_i = 1) > \pi_j \Pr(\mathbf{z}_{\mathcal{A}}|X_j = 1)\ \&\ (h_i \neq h_j) \iff h_j > h_i \quad (8a)$$

$$\text{and } \pi_i \overline{\Pr}(\mathbf{z}_{\mathcal{A}}|X_i = 1) > \pi_j \overline{\Pr}(\mathbf{z}_{\mathcal{A}}|X_j = 1)\ \&\ (h_i \neq h_j) \iff h_j > h_i. \quad (8b)$$

We will first prove (8a), followed by (8b). Note that $h_j > h_i$ is equivalent to $h_j \geq h_i + 1$. Using the fact that $p < \frac{\rho}{1+\rho}$ and that for any i, j, $\frac{\pi_j}{\pi_i} \leq \frac{\max_k \pi_k}{\min_k \pi_k} = \frac{1}{\rho}$, we can show the converse of (8a) as follows. If $h_j - h_i \geq 1$, then

$$\begin{aligned}
&(h_j - h_i)\log\frac{1-p}{p} \geq \log\frac{1-p}{p} > \log\frac{1}{\rho} \geq \log\frac{\pi_j}{\pi_i} \\
&\implies \log\pi_i + h_i\log\frac{p}{1-p} > \log\pi_j + h_j\log\frac{p}{1-p} \\
&\implies \log\pi_i p^{h_i}(1-p)^{k-h_i} > \log\pi_j p^{h_j}(1-p)^{k-h_j}.
\end{aligned}$$

To prove the forward direction, we need to show that

$$h_j \leq h_i \implies (h_i = h_j) \text{ or } \pi_i \Pr(\mathbf{z}_{\mathcal{A}}|X_i = 1) \leq \pi_j \Pr(\mathbf{z}_{\mathcal{A}}|X_j = 1).$$

If $h_j < h_i$, then $\pi_i \Pr(\mathbf{z}_{\mathcal{A}}|X_i = 1) < \pi_j \Pr(\mathbf{z}_{\mathcal{A}}|X_j = 1)$ using the converse result with dummy variables i and j interchanged, thereby proving (8a). Similarly, to prove the converse of (8b), we need to show that $h_j > h_i$ leads to $\pi_i \overline{\Pr}(\mathbf{z}_{\mathcal{A}}|X_i = 1) > \pi_j \overline{\Pr}(\mathbf{z}_{\mathcal{A}}|X_j = 1)$, for which we need to consider three different cases.

Case 1 : Let $h_j > h_i \geq k\overline{p} \implies \varepsilon_i = \varepsilon_j = \overline{p}$. Then,

$$\begin{aligned}
&(h_j - h_i)\log\frac{1-\overline{p}}{\overline{p}} \geq \log\frac{1-\overline{p}}{\overline{p}} > \log\frac{1}{\rho} \geq \log\frac{\pi_j}{\pi_i} \\
&\implies \log\pi_i + h_i\log\frac{\overline{p}}{1-\overline{p}} > \log\pi_j + h_j\log\frac{\overline{p}}{1-\overline{p}} \\
&\implies \log\pi_i \overline{p}^{h_i}(1-\overline{p})^{k-h_i} > \log\pi_j \overline{p}^{h_j}(1-\overline{p})^{k-h_j} \\
&\implies \log\pi_i \varepsilon_i^{h_i}(1-\varepsilon_i)^{k-h_i} > \log\pi_j \varepsilon_j^{h_j}(1-\varepsilon_j)^{k-h_j}.
\end{aligned}$$

Case 2 : Let $h_j \geq k\overline{p} > h_i \implies \varepsilon_i = h_i/k$ and $\varepsilon_j = \overline{p}$. Then, following along the same lines as above, we have

[2] available at `www-personal.umich.edu/~gowtham`

$$\log \pi_i \overline{p}^{h_i}(1-\overline{p})^{k-h_i} > \log \pi_j \overline{p}^{h_j}(1-\overline{p})^{k-h_j}$$
$$\implies \log \pi_i \left(\frac{h_i}{k}\right)^{h_i}\left(1-\frac{h_i}{k}\right)^{k-h_i} > \log \pi_j \overline{p}^{h_j}(1-\overline{p})^{k-h_j}$$
$$\implies \log \pi_i \varepsilon_i^{h_i}(1-\varepsilon_i)^{k-h_i} > \log \pi_j \varepsilon_j^{h_j}(1-\varepsilon_j)^{k-h_j}$$

where the second statement follows from (4) in Lemma 1. Case 3 : Let $k\overline{p} > h_j > h_i$, which implies $\varepsilon_i = h_i/k$ and $\varepsilon_j = h_j/k$. Defining $g_1(h) = \log[(h/k)^h(1-h/k)^{k-h}]$ and $g_2(h) = \log \overline{p}^h(1-\overline{p})^{k-h}$, we have,

$$\frac{dg_1}{dh} = \log\frac{h/k}{1-\frac{h}{k}} < \frac{dg_2}{dh} = \log\frac{\overline{p}}{1-\overline{p}} < 0,$$

when $h < k\overline{p}$. This implies that $g_1(h)$ has a larger slope than $g_2(h)$ when $h \in [0, k\overline{p})$, and hence

$$\log(\varepsilon_i)^{h_i}(1-\varepsilon_i)^{k-h_i} - \log(\varepsilon_j)^{h_j}(1-\varepsilon_j)^{k-h_j}$$
$$> \log \overline{p}^{h_i}(1-\overline{p})^{k-h_i} - \log \overline{p}^{h_j}(1-\overline{p})^{k-h_j}$$
$$= (h_j - h_i)\log\frac{1-\overline{p}}{\overline{p}} > \log\frac{\pi_j}{\pi_i}$$
$$\implies \log \pi_i \varepsilon_i^{h_i}(1-\varepsilon_i)^{k-h_i} > \log \pi_j \varepsilon_j^{h_j}(1-\varepsilon_j)^{k-h_j},$$

thus proving the converse of (8b). The forward direction can be proved using the converse result in the same way as it is done for (8a). □

5.2 Response-Dependent Noise

We now consider the noise model where the probability of error depends on the true response. When the true response is 0, the probability of observing a noisy response is given by ν_0, and by ν_1 when the true response is 1, i.e.,

$$\Pr(Z_j = 0|X_i = 1) = 1-\nu_0, \text{ if } b_{ij} = 0,$$
$$\text{and } \Pr(Z_j = 0|X_i = 1) = \nu_1, \text{ if } b_{ij} = 1.$$

For example, consider the noise model where $\gamma_{ij}^0 = 0$, $\gamma_{ij}^1 = \gamma$, $\forall i,j$, and $0 < p_j^0 = p^0 < 1$, $0 < p_j^1 = p^1 < 1$, $\forall j$. The probability of error depends only on the true response with $\nu_0 = p^0$ and $\nu_1 = (1-\gamma)p^1 + \gamma(1-p^0)$.

Similarly, the noise model in the QMR-DT problem in the case of single fault can be reduced to this setting with $\nu_0 = \rho_l$ and $\nu_1 = (1-\rho_l)\rho_i$, where $0 < \rho_l, \rho_i < 1$ are referred to as the leak probability and inhibition probability, respectively (Zheng et al., 2005). This noise model is often used in the context of fault diagnosis, and is also a special case of the general model in Section 2.

For any subset of indices $\mathcal{A} \subseteq \{1, \cdots, N\}$, let $\mathcal{A}_0^i = \{j \in \mathcal{A} : b_{ij} = 0\}$ and $\mathcal{A}_1^i = \{j \in \mathcal{A} : b_{ij} = 1\}$ be partitions of $\mathcal{A}$ for each $i = 1, \cdots, M$ such that the true response b_{ij} of object θ_i to queries in $Q_{\mathcal{A}_0^i}$ is 0, and that in $Q_{\mathcal{A}_1^i}$ is 1. Then, the likelihood function is given by

$$\Pr(\mathbf{Z}_\mathcal{A} = \mathbf{z}_\mathcal{A}|X_i = 1) = \nu_0^{\delta_{i,\mathcal{A}_0^i}}(1-\nu_0)^{|\mathcal{A}_0^i|-\delta_{i,\mathcal{A}_0^i}} \cdot \nu_1^{\delta_{i,\mathcal{A}_1^i}}(1-\nu_1)^{|\mathcal{A}_1^i|-\delta_{i,\mathcal{A}_1^i}}$$

where $\delta_{i,\mathcal{A}_0^i} = \sum_{j\in\mathcal{A}_0^i}|0-z_j|$ and $\delta_{i,\mathcal{A}_1^i} = \sum_{j\in\mathcal{A}_1^i}|1-z_j|$, are local Hamming distances between the true responses of θ_i to queries in $Q_{\mathcal{A}_0^i}$ and $Q_{\mathcal{A}_1^i}$, and that of their observed responses. Once again, using Lemma 1, this likelihood function can be upper bounded by

$$\overline{\Pr}(\mathbf{Z}_\mathcal{A} = \mathbf{z}_\mathcal{A}|X_i = 1) = \left(1-\frac{\delta_{i,\mathcal{A}_0^i}}{|\mathcal{A}_0^i|}\right)^{|\mathcal{A}_0^i|-\delta_{i,\mathcal{A}_0^i}} \left(\frac{\delta_{i,\mathcal{A}_0^i}}{|\mathcal{A}_0^i|}\right)^{\delta_{i,\mathcal{A}_0^i}} \cdot \left(1-\frac{\delta_{i,\mathcal{A}_1^i}}{|\mathcal{A}_1^i|}\right)^{|\mathcal{A}_1^i|-\delta_{i,\mathcal{A}_1^i}} \left(\frac{\delta_{i,\mathcal{A}_1^i}}{|\mathcal{A}_1^i|}\right)^{\delta_{i,\mathcal{A}_1^i}}.$$

Hence, the ranks of the objects can be estimated using (5) and the rank-based query selection can be performed using (6), without requiring any knowledge of the query noise parameters.

Unfortunately, it is not possible to extend the result of Proposition 1 to this case. Yet, the experimental results in Section 6 demonstrate that the noise-independent rank-based algorithm performs comparably to the entropy-based algorithm, which requires knowledge of ν_0 and ν_1.

6 Experiments

We compare the performance of the proposed rank-based algorithm with entropy-based query selection, GBS, and random search, on 2 synthetic datasets, 1 semi-synthetic dataset, and 1 real dataset. GBS and random search serve as baselines and are not expected to perform well since GBS doesn't account for noise, and random search just selects queries at random.

The first two datasets are random bipartite networks (Guillaume and Latapy, 2004) generated using the standard Erdös-Rényi (ER) random network model and the Preferential Attachment (PA) random network model. The third dataset is a network topology built using the BRITE generator (Medina et al., 2001), which simulates an Internet-like topology at the Autonomous Systems level. To generate a bipartite network of components and probes from the BRITE network, we used the approach described by Rish et al. (2005) and Zheng et al. (2005). The last dataset is the WISER database, which is a toxic chemical database describing the binary relation between 298 toxic chemicals and 79 acute symptoms (Szczur and Mashayekhi, 2005).

We generated a random network for each of the random network models considered, where each network consisted of around 200 objects and 300 queries. We generated a BRITE network consisting of 300 objects (components/computers) and around 350 queries (probes). For the synthetic datasets and WISER, we assumed the noise

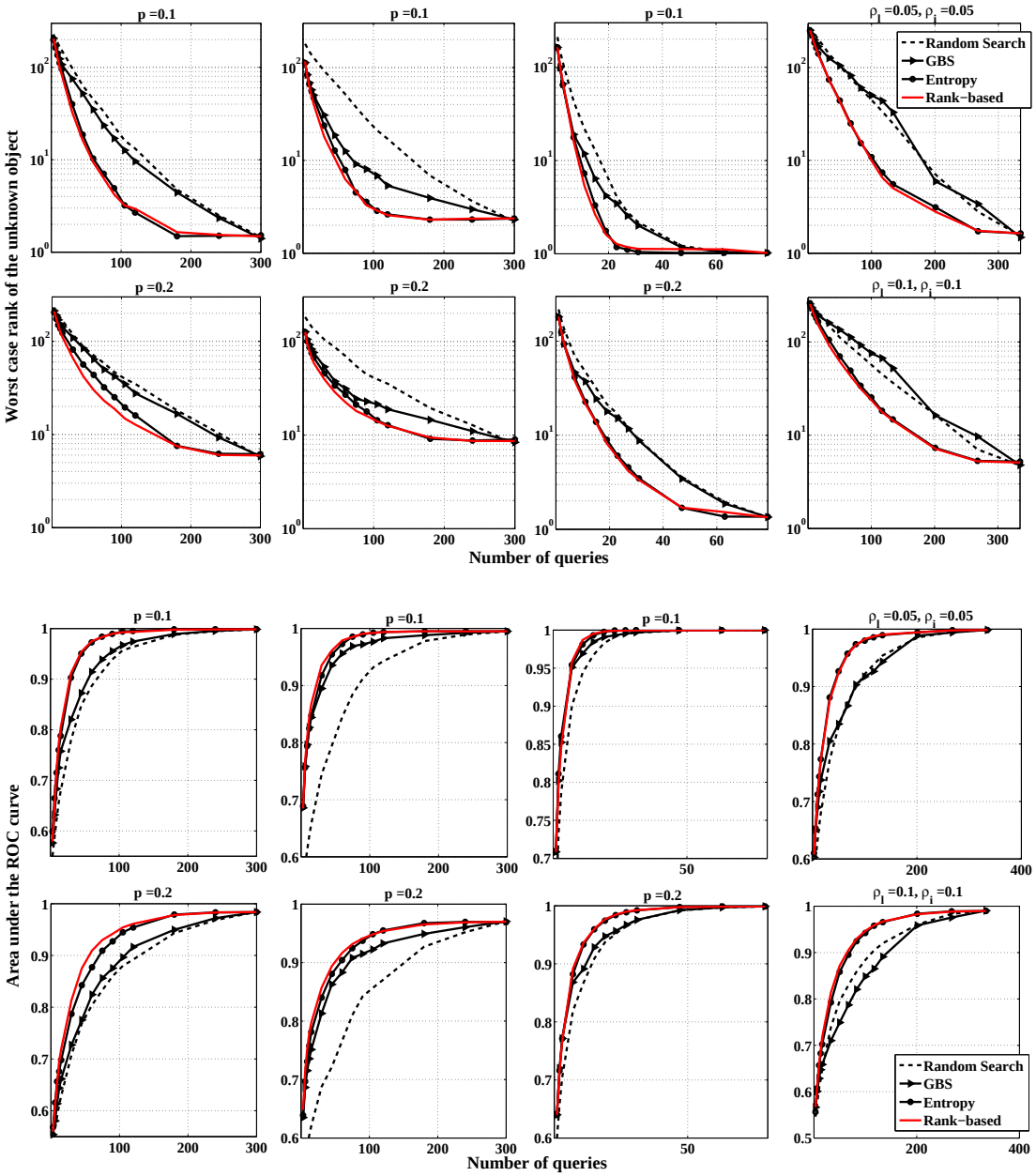

Figure 4: The plots in the first column correspond to a dataset generated using ER model, the second column correspond to a dataset generated using the PA model, the third column corresponds to the WISER database and the last column corresponds to a BRITE network. In all the experiments, the rank-based algorithm is performed without knowledge of the noise parameters. These experiments also demonstrate that θ_{MAP} is not necessarily equal to the unknown object θ.

model to be that of Section 5.1, and for the BRITE network, we considered the noise model in Section 5.2. Here, we present the results under uniform prior where $\pi_i = 1/M$. We observed similar performance under non-uniform prior as shown in the Supplementary material.

Figure 4 shows the worst case rank of the unknown object θ and the area under the ROC curve as a function of the number of queries inputted. The ROC curve is generated as follows: After observing responses to a set of queries, the objects are ranked based on their posterior probabilities where ties involving objects with equal posterior probabilities are broken randomly, instead of a worst case ranking. Given such a ranking of the objects in Θ, the ROC curve can be obtained by varying the threshold t, where the states of the top t objects are declared as 1 and the rest 0 leading to a certain miss rate and false alarm rate. Refer to Supplementary material for more details.

Each curve in these figures is averaged over 500 random

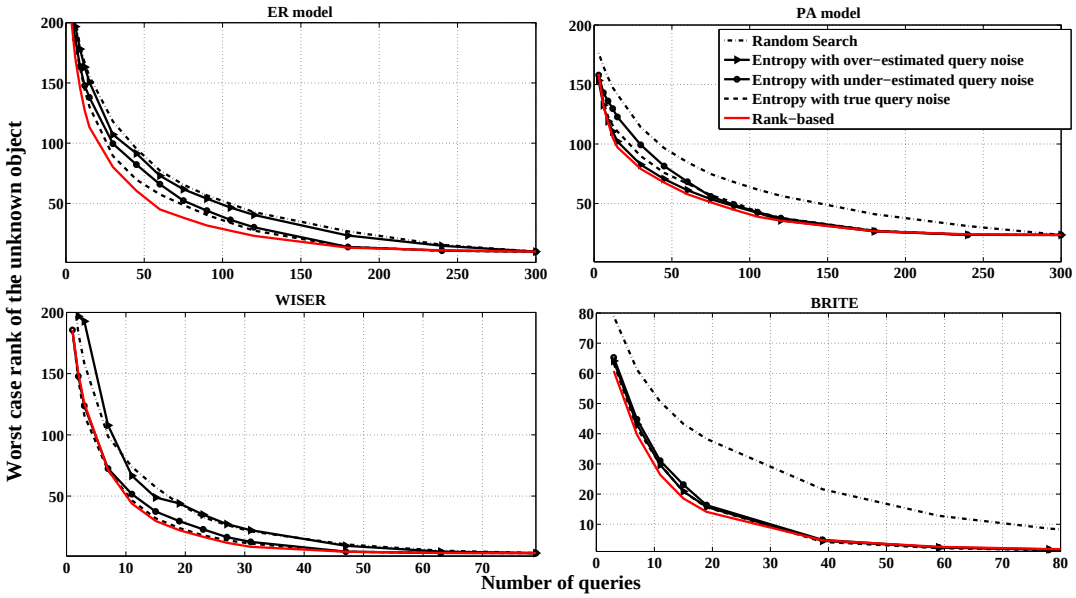

Figure 5: Demonstrates the sensitivity of entropy-based query selection to mis-specification of noise parameters

realizations, where each random realization corresponds to a random selection of $\theta \in \Theta$ and random generation of the noisy query responses. The plots in the first column correspond to a dataset generated using the ER model, the second column corresponds to the PA model, the third column corresponds to the WISER database, and the last column to a BRITE network. For the 2 random network models and BRITE, the results were observed to be consistent across different realizations of the underlying bipartite network.

For the ER, PA, and the WISER datasets, we consider two different values for the probability of error, $p = 0.1, 0.2$. The entropy-based query selection is performed assuming the knowledge of p, whereas the rank-based query selection is performed using only the fact that $p < \overline{p} = 0.5$. The BRITE networks are simulated using the QMR-DT noise model, where we considered the leak and the inhibition probabilities to be $(\rho_i, \rho_l) = (0.05, 0.05)$ and $(0.1, 0.1)$. This noise model reduces to that in Section 5.2 with $\nu_0 = \rho_l$ and $\nu_1 = (1-\rho_l)\rho_i$. Once again, the entropy-based query selection is performed assuming the knowledge of ν_0 and ν_1, whereas the rank-based query selection is performed using only the fact that $\nu_0, \nu_1 \leq \overline{p} = 0.25$. Also, note from these plots that θ_{MAP} is not always equal to the unknown object θ.

Finally, Figure 5 demonstrates the sensitivity of entropy-based query selection to mis-specification of the value of noise parameters. For the ER, PA and the WISER datasets, the true noise parameter is $p = 0.25$ while the under-estimated and the over-estimated curves are obtained using $p = 0.15$ and 0.4, respectively. For the BRITE network, while the true noise parameters are $(0.1, 0.1)$, the other two curves are obtained using $(0.05, 0.05)$ and $(0.15, 0.15)$. Once again, the rank-based algorithm is performed without knowledge of the noise parameters. This demonstrates that the entropy-based query selection can perform poorly when the noise parameters are mis-specified.

These experiments demonstrate the competitive performance of the rank-based algorithm to entropy-based query selection, despite not having the knowledge of the underlying noise parameters.

7 Conclusions

We study the problem of active diagnosis under persistent noise, and propose a rank-based greedy algorithm. In this algorithm, queries are selected sequentially such that the expected rank of the unknown object is minimized, and the output is a ranked list of the objects rather than the object θ_{MAP} with the maximum *a posteriori* probability. Unlike traditional approaches such as mutual information (or conditional entropy) based query selection, the rank-based algorithm does not require the knowledge of the underlying query noise. In addition, we show that in certain noise models, the ranks estimated with limited knowledge of the noise parameters are equal to the true ranks. Finally, we demonstrate through experiments on real and synthetic datasets, the competitive performance of the proposed algorithm to entropy-based query selection, despite not having the knowledge of the underlying query noise.

Acknowledgments

The authors would like to thank the anonymous reviewers and Jason Stanley for providing constructive feedback. G. Bellala and C. Scott were supported in part by NSF Awards No. 0830490 and 0953135. All the authors were supported in part by CDC/NIOSH grant No. R21 OH009441-01A2.

References

M. F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

G. Bellala, S. K. Bhavnani, and C. Scott. Group-based query learning for rapid diagnosis in time-critical situations. Technical report, 2009. available online at arXiv.org:0911.4511.

S. K. Bhavnani, A. Abraham, C. Demeniuk, M. Gebrekristos, A. Gong, S. Nainwal, G. K. Vallabha, and R. J. Richardson. Network analysis of toxic chemicals and symptoms: Implications for designing first-responder systems. *Proceedings of American Medical Informatics Association (AMIA)*, 2007.

S. Dasgupta. Analysis of a greedy active learning strategy. *Advances in Neural Information Processing Systems (NIPS)*, 2004.

M. Garey. Optimal binary identification procedures. *SIAM Journal on Applied Mathematics*, 23(2):173–186, 1972.

D. Geman and B. Jedynak. An active testing model for tracking roads in satellite images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1):1–14, 1996.

D. Golovin and A. Krause. Adaptive Submodularity: A new approach to active learning and stochastic optimization. *In Proceedings of International Conference on Learning Theory (COLT)*, 2010.

D. Golovin, D. Ray, and A. Krause. Near-optimal Bayesian active learning with noisy observations. *to appear in the Proceedings of the Neural Information Processing Systems (NIPS)*, 2010.

J. Guillaume and M. Latapy. *Bipartite Graphs as Models of Complex Networks*. Springer, 2004.

A. Gupta, R. Krishnaswamy, V. Nagarajan, and R. Ravi. Approximation algorithms for optimal decision trees and adaptive TSP problems. CoRR, abs/1003.0722, 2010.

S. Hanneke. Teaching dimension and the complexity of active learning. *Proceedings of the 20th Conference on Learning Theory (COLT)*, 2007.

L. Hyafil and R. Rivest. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1):15–17, 1976.

M. Kääriäinen. Active learning in the non-realizable case. *Algorithmic Learning Theory*, pages 63–77, 2006.

A. P. Korostelev and J. C. Kim. Rates of convergence of the sup-norm risk in image models under sequential designs. *Statistics and Probability letters*, 46:391–399, 2000.

S. R. Kosaraju, T. M. Przytycka, and R. S. Borgstrom. On an optimal split tree problem. *Proceedings of 6th International Workshop on Algorithms and Data Structures, WADS*, pages 11–14, 1999.

T. Le and C. N. Hadjicostis. Max-product algorithms for the generalized multiple-fault diagnosis problem. *IEEE Transactions on Systems, Man and Cybernetics*, 37(6), December 2007.

D. W. Loveland. Performance bounds for binary testing with arbitrary weights. *Acta Informatica*, 1985.

A. Medina, A. Lakhina, I. Matta, and J. Byers. BRITE: An Approach to Universal Topology Generation. *In Proceedings of the International Workshop on Modeling, Analysis and Simulation of Computer and Telecommunications Systems (MASCOT)*, 2001.

R. Nowak. Generalized binary search. *Proceedings of the 46th Allerton Conference on Communications, Control and Computing*, pages 568–574, 2008.

R. Nowak. Noisy generalized binary search. *Advances in Neural Information Processing Systems (NIPS)*, 22: 1366–1374, 2009.

A. Rényi. On a problem of information theory. *MTA Mat. Kut. Int. Kozl.*, 6B:505 – 516, 1961.

I. Rish, M. Brodie, S. Ma, N. Odintsova, A. Beygelzimer, G. Grabarnik, and K. Hernandez. Adaptive diagnosis in distributed systems. *IEEE Transactions on Neural Networks*, 16(5):1088 – 1109, 2005.

M. J. Swain and M. A. Stricker. Promising directions in active vision. *International Journal of Computer Vision*, 11(2):109–126, 1993.

M. Szczur and B. Mashayekhi. WISER Wireless Information System for Emergency Responders. *Proceedings of American Medical Informatics Association Annual Symposium*, 2005.

S. Yu, B. Krishnapuram, R. Rosales, and R. B. Rao. Active sensing. *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

A. X. Zheng, I. Rish, and A. Beygelzimer. Efficient test selection in active diagnosis via entropy approximation. *In Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2005.