Electronic Journal of Statistics
Vol. 11 (2017) 4969–5017
ISSN: 1935-7524
https://doi.org/10.1214/17-EJS1336SI

Near-optimal Bayesian active learning with correlated and noisy tests

Yuxin Chen

California Institute of Technology
e-mail: chenyux@caltech.edu

S. Hamed Hassani

University of Pennsylvania
e-mail: hassani@seas.upenn.edu

and

Andreas Krause

ETH Zurich
e-mail: krausea@ethz.ch

Abstract: We consider the Bayesian active learning and experimental design problem, where the goal is to learn the value of some unknown target variable through a sequence of informative, noisy tests. In contrast to prior work, we focus on the challenging, yet practically relevant setting where test outcomes can be conditionally *dependent* given the hidden target variable. Under such assumptions, common heuristics, such as greedily performing tests that maximize the reduction in uncertainty of the target, often perform poorly.

We propose ECED, a novel, efficient active learning algorithm, and prove strong theoretical guarantees that hold with correlated, noisy tests. Rather than directly optimizing the prediction error, at each step, ECED picks the test that maximizes the gain in a surrogate objective, which takes into account the dependencies between tests. Our analysis relies on an information-theoretic auxiliary function to track the progress of ECED, and utilizes adaptive submodularity to attain the approximation bound. We demonstrate strong empirical performance of ECED on three problem instances, including a Bayesian experimental design task intended to distinguish among economic theories of how people make risky decisions, an active preference learning task via pairwise comparisons, and a third application on pool-based active learning.

Keywords and phrases: Bayesian active learning, information gathering, decision making, noisy observation, approximation algorithms.

Received June 2017.

Contents

1. Introduction

Optimal information gathering, i.e., selectively acquiring the most useful data, is one of the central challenges in interactive machine learning. The problem of optimal information gathering has been studied in the context of active instance labeling [9, 28], active feature evaluation[1] [21, 11, 9, 28], Bayesian experimental design [12, 5], policy making [17, 27], probabilistic planning and optimal control [31], and numerous other domains. In a typical set-up for these problems, there is some unknown *target variable* Y of interest, and a set of *tests*, which correspond to observable variables defined through a probabilistic model. The goal

[1] Structurally, the problem of active feature evaluation is the same with active instance labeling, and hence the term "Bayesian active learning" is used to refer to both cases.

is to determine the value of the target variable via a sequential *policy*, which adaptively selects the next test based on previous observations, such that the cost of performing these tests is minimized.

Deriving the optimal testing policy is NP-hard in general [4]; however, under certain conditions, some approximation results are known. In particular, if test outcomes are deterministic functions of the target variable (i.e., in the *noise-free* setting), a simple greedy algorithm, namely Generalized Binary Search (GBS), is guaranteed to provide a near-optimal approximation of the optimal policy [23]. On the other hand, if test outcomes are noisy, but the outcomes of different tests are *conditionally independent* given Y (i.e., under the Naïve Bayes assumption), then using the most informative selection policy, which greedily selects the test that maximizes the expected reduction in uncertainty of the target variable (quantified in terms of Shannon entropy), is guaranteed to perform near-optimally [7].

However, in many practical problems, due to the effect of noise or complex structural dependencies in the probabilistic model (beyond Naïve Bayes), we only have access to tests that are *indirectly informative* about the target variable Y (i.e., test outcomes depend on Y through another hidden random variable. See Fig. 1.) – as a consequence, the test outcomes become conditionally dependent given Y. Consider a medical diagnosis example, where a doctor wants to predict the best treatment for a patient, by carrying out a series of medical tests, each of which reveals some information about the patient's physical condition. Here, outcomes of medical tests are conditionally independent given the patient's condition, but are *not* independent given the treatment, which is made based on the patient's condition. It is known that in such cases, both GBS and the most informative selection policy (which myopically maximizes the information gain w.r.t. the distribution over Y) can perform arbitrarily poorly. Golovin et al. [14] then formalize this problem as an *equivalence class determination* problem (See §2.3), and show that if the tests' outcomes are noise-free, then one can obtain near-optimal expected cost, by running a greedy policy based on a surrogate objective function. Their results rely on the fact that the surrogate objective function exhibits *adaptive submodularity* [13], a natural diminishing returns property that generalizes the classical notion of submodularity to adaptive policies. Unfortunately, in the general setting where tests are noisy, no efficient policies are known to be provably competitive with the optimal policy.

Our contribution. In this paper, we introduce *Equivalence Class Edge Discounting (*ECED*)*, a novel algorithm for practical Bayesian active learning and experimental design problems, and prove strong theoretical guarantees with correlated, noisy tests. In particular, we focus on the setting where the tests' outcomes indirectly depend on the target variable (and hence are conditionally dependent given Y), and we assume that the outcome of each test can be corrupted by some random, *persistent* noise[2] (§2). We prove that when the test outcomes are binary, and the noise on test outcomes are mutually indepen-

[2]Persistent noise means that repeating a test produces identical outcomes.

dent, then ECED is guaranteed to obtain near-optimal cost, compared with an optimal policy that achieves a lower prediction error (§3). We develop a theoretical framework for analyzing such sequential policies, where we leverage an information-theoretic auxiliary function to reason about the effect of noise, and combine it with the theory of adaptive submodularity to attain the approximation bound (§4). The key insight is to show that ECED is effectively making progress in the long run as it picks more tests, even if the myopic choices of tests do not have immediate gain in terms of reducing the uncertainty of the target variable. We demonstrate the compelling performance of ECED on two real-world problem instances: A Bayesian experimental design task intended to distinguish among economic theories of how people make risky decisions, an active preference learning task via pairwise comparisons and a third application on pool-based active learning (§5). To facilitate better understanding, we provide illustrative examples and full proofs of our theoretical results in the Appendix.

2. Preliminaries and problem statement

2.1. The basic model

Let Y be the target random variable whose value we want to learn. The value of Y, which ranges among set $\mathcal{Y} = \{y_1, \ldots, y_t\}$, depends deterministically on another random variable $\Theta \in \text{supp}(\Theta) = \{\theta_1, \ldots, \theta_n\}$ with some known distribution $\mathbb{P}[\Theta]$. Concretely, there is a deterministic mapping $r : \text{supp}(\Theta) \to \mathcal{Y}$ that gives $Y = r(\Theta)$ (see Fig. 1).

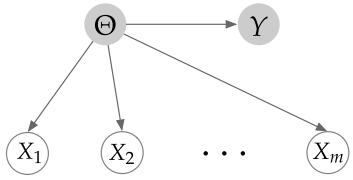


FIG 1. *The basic model*

Let $\mathcal{X} = \{X_1, \ldots, X_m\}$ be a collection of discrete observable variables that are statistically dependent on Θ. We use $e \in \mathcal{V} \triangleq \{1, \ldots, m\}$ as the indexing variable of a test. Performing each test X_e produces an outcome $x_e \in \mathcal{O}$ (here, $\mathcal{O}$ encodes the set of possible outcomes of a test), and incurs a *unit cost*. We can think of Θ as representing the underlying "root-cause" among a set of n possible root-causes of the joint event $\{X_1, \ldots, X_m\}$, and Y as representing the optimal "target action" to be taken for root-cause Θ. Also, each of the X_e's is a "test" that we can perform, whose observation reveals some information about Θ. In our medical diagnosis example (see Fig. 2(a)), X_e's encode tests' outcomes, Y encodes the treatment, and Θ encodes the patient's physical condition.

Crucially, we assume that X_e's are *conditionally independent*[3] given Θ, i.e., $\mathbb{P}[\Theta, X_1, \ldots, X_m] = \mathbb{P}[\Theta] \prod_{i=1}^{m} \mathbb{P}[X_i \mid \Theta]$ with known parameters. Note that noise is implicitly encoded in our model, as we can equivalently assume that X_e's are first generated from a deterministic mapping of Θ, and then perturbed by some random noise. As an example, if test outcomes are binary, then we can think of X_e as resulting from flipping the deterministic outcome of test e given Θ with some probability, and the flipping events of the tests are mutually independent.

2.2. Problem statement

We consider sequential, adaptive policies for picking the tests. Denote a policy by π. In words, a policy specifies which test to pick next, as well as when to stop picking tests, based on the tests picked so far and their corresponding outcomes. After each pick, our observations so far can be represented as a partial realization $\Psi \in 2^{\mathcal{V}\times\mathcal{O}}$ (e.g., Ψ encodes what tests have been performed and what their outcomes are). Formally, a policy $\pi : 2^{\mathcal{V}\times\mathcal{O}} \nrightarrow \mathcal{V}$ is defined to be a partial mapping from partial realizations Ψ to tests.

Suppose that running π till termination returns a sequence of test-observation pairs of length k, denoted by ψ_π, i.e.,

$$\psi_\pi \triangleq \{(e_{\pi,1}, x_{e_{\pi,1}}), (e_{\pi,2}, x_{e_{\pi,2}}), \cdots, (e_{\pi,k}, x_{e_{\pi,k}})\}.$$

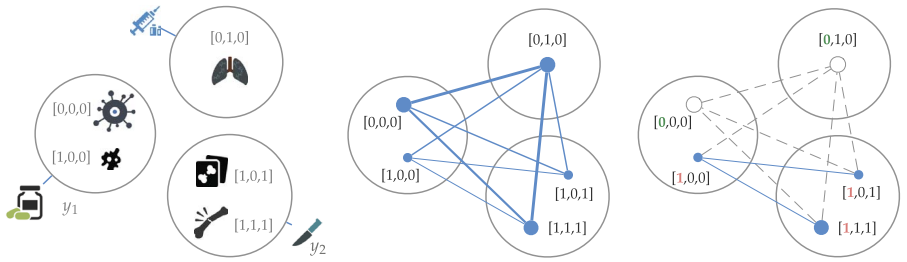
This can be interpreted as a random path taken by policy π. Once ψ_π is observed, we obtain a new posterior on Θ (and consequently on Y). The best prediction one can thus make under the Bayesian setting is the MAP estimator $\hat{y}$ of Y, i.e., $\hat{y} \triangleq \arg\max_{y' \in \mathcal{Y}} \mathbb{P}[Y = y' \mid \psi_\pi]$. The error probability of predicting $\hat{y}$ is

$$p_{\mathrm{ERR}}^{\mathrm{MAP}}(\psi_\pi) \triangleq \mathbb{P}[\hat{y} \neq y \mid \psi_\pi] = 1 - \max_{y \in \mathcal{Y}} \mathbb{P}[y \mid \psi_\pi].$$

We call $p_{\mathrm{ERR}}^{\mathrm{MAP}}$ the *prediction error* of the MAP estimator. The expected prediction error after running policy π is then defined as $p_{\mathrm{ERR}}(\pi) \triangleq \mathbb{E}_{\psi_\pi}\left[p_{\mathrm{ERR}}^{\mathrm{MAP}}(\psi_\pi)\right]$. Let the (worst-case) cost of π be $\mathrm{cost}(\pi) \triangleq \max_{\psi_\pi} |\psi_\pi|$, i.e., the maximum number of tests performed by π over all possible paths it takes. Given some small tolerance $\delta \in [0, 1]$, we seek a policy with the minimal cost, such that upon termination, the posterior puts at least $1 - \delta$ mass on the most likely target value y in expectation. In other words, we require that the expected prediction error after running the policy is at most δ. Denote such policy by $\mathsf{OPT}(\delta)$. Formally, we seek

$$\mathsf{OPT}(\delta) \in \arg\min_{\pi} \mathrm{cost}(\pi), \text{ s.t. } p_{\mathrm{ERR}}(\pi) \leq \delta. \tag{2.1}$$

[3] In active instance selection, this simply implies that labeling errors are independent, which is a standard assumption made in the statistical learning literature.



(a) Medical diagnosis example (b) Initialization (c) EC^2

FIG 2. *(a) shows an illustrative example of the medical diagnosis problem: there are 5 root-causes with different symptoms represented by the 3-dimensional binary vectors; each of the circles represents a medical treatment for the root-causes enclosed. In (b), we initialize* EC^2*, by drawing edges between all pairs of root-causes (solid dots) that are mapped into different treatments (circles). In (c), we run* EC^2 *and remove all the edges incident to root-causes* $\theta_2[0,0,0]$ *and* $\theta_5[0,1,0]$ *if we observe* $X_1 = 1$.

Remarks. Note that there are different ways of defining "success" of a policy. Other than bounding the prediction error as considered in Eq. (2.1), an alternative option is to ensure that the *excess error*, or *regret* of acting upon ψ_π, compared to having observed all the tests is not more than δ. While the regret-based success criterion might be an alternative sensible criterion to consider, the prediction error criterion offers a natural stopping condition for running a policy (as one can compute the $p_{\mathrm{ERR}}^{\mathrm{MAP}}(\psi_\pi)$ purely based on the posterior). Hence we focus on Problem 2.1 throughout this paper.

2.3. Special case: The equivalence class determination problem

Computing the optimal policy for Problem (2.1) is intractable in general. When $\delta = 0$, this problem reduces to the equivalence class determination problem [14, 2]. Here, the target variables are referred to as *equivalence classes*, since each $y \in \mathcal{Y}$ corresponds to a subset of root-causes in $\mathrm{supp}(\Theta)$ that (equivalently) share the same "action".

Noise-free setting: The EC^2 algorithm. Let us first assume that tests are noise-free, i.e., $\forall e,\ \mathbb{P}[X_e \mid \Theta] \in \{0,1\}$. Then this problem can be solved near-optimally by the *equivalence class edge cutting* (EC^2) algorithm [14]. As illustrated in Fig. 2, EC^2 employs an edge-cutting strategy based on a weighted graph $G = (\mathrm{supp}(\Theta), E)$, where vertices represent root-causes, and edges link root-causes that we want to distinguish between. Formally, $E \triangleq \{(\theta,\theta') : r(\theta) \neq r(\theta')\}$ consists of all (unordered) pairs of root-causes corresponding to different target values (see Fig. 2(b)). We define a weight function $w : E \to \mathbb{R}_{\geq 0}$ by $w((\theta,\theta')) \triangleq \mathbb{P}[\theta] \cdot \mathbb{P}[\theta']$, i.e., as the product of the probabilities of its incident root-causes. We extend the weight function on sets of edges $E' \subseteq E$ as the sum of weight of all edges $(\theta,\theta') \in E'$, i.e., $w(E') \triangleq \sum_{(\theta,\theta')\in E'} w((\theta,\theta'))$.

Performing test $e \in \mathcal{V}$ with outcome x_e is said to "*cut*" an edge, if at least one of its incident root-causes is inconsistent with x_e (See Fig. 2(c)). Denote $E(x_e) \triangleq \{\{\theta, \theta'\} \in E : \mathbb{P}[x_e \mid \theta] = 0 \ \vee \ \mathbb{P}[x_e \mid \theta'] = 0\}$ as the set of edges cut by observing x_e. The EC^2 objective (which is greedily maximized per iteration of EC^2), is then defined as the total weight of edges cut by the current partial observation ψ_π: $f_{\mathrm{EC}^2}(\psi_\pi) \triangleq w\Big(\bigcup_{(e,x_e)\in\psi_\pi} E(x_e)\Big)$.

The EC^2 objective function is *adaptive submodular*, and *strongly adaptive monotone*. Formally, let $\psi_1, \psi_2 \in 2^{\mathcal{V}\times\mathcal{O}}$ be two partial realizations of tests' outcomes. We call ψ_1 a *subrealization* of ψ_2, denoted as $\psi_1 \preceq \psi_2$, if every test seen by ψ_1 is also seen by ψ_2, and $\mathbb{P}[\psi_2 \mid \psi_1] > 0$. A function $f : 2^{\mathcal{V}\times\mathcal{O}} \to \mathbb{R}$ is called *adaptive submodular* w.r.t. a distribution $\mathbb{P}$, if for any $\psi_1 \preceq \psi_2$ and any X_e it holds that $\Delta(X_e \mid \psi_1) \geq \Delta(X_e \mid \psi_2)$, where $\Delta(X_e \mid \psi) := \mathbb{E}_{x_e}[f(\psi \cup \{(e, x_e)\}) - f(\psi) \mid \psi]$ (i.e., "adding information earlier helps more"). Further, function f is called *strongly adaptively monotone* w.r.t. $\mathbb{P}$, if for all ψ, test e not seen by ψ, and $x_e \in \mathcal{O}$, it holds that $f(\psi) \leq f(\psi \cup \{(e, x_e)\})$ (i.e., "adding new information never hurts"). For sequential decision problems satisfying adaptive submodularity and strongly adaptive monotonicity, the policy that greedily, upon having observed ψ, selects the test $e^* \in \arg\max_e \Delta(X_e \mid \psi)$, is guaranteed to attain near-minimal cost [13].

Noisy setting. Notice that, the EC^2 algorithm can, to some extent, deal with noisy observations. In particular, for noise with "small" support (e.g., assume that for any root-cause Θ, a maximal number of k tests are allowed to be corrupted, where k is some finite integer), one can reduce the noisy problem to a noiseless one, by enumerating all possible realizations of tests, and treat each realization as a new "root-cause". However, for the more general setting with *i.i.d.* noise (e.g., $\mathbb{P}[X_e \mid \Theta] \in (0, 1)$), it may not be possible to cut all the edges constructed for EC^2 (or equivalently, to attain 0 error probability in prediction Y), even if we exhaust all tests. Hence the theoretical results of Golovin et al. [14] no longer apply. A natural approach to solving Problem (2.1) for $\delta > 0$ would be to pick tests greedily maximizing the expected reduction in the error probability p_{ERR}. However, this objective is not adaptive submodular; in fact, as we show in Appendix C, such policy can perform arbitrarily badly if there are complementaries among tests, i.e., the gain of a set of tests can be far better than sum of the individual gains of the tests in the set. Therefore, motivated by the EC^2 objective in the noise-free setting, we would like to optimize a surrogate objective function which captures the effect of noise, while being amenable to greedy optimization.

3. The ECED algorithm

We now introduce ECED for Bayesian active learning under correlated noisy tests, which strictly generalizes EC^2 to the noisy setting, while preserving the near-optimal guarantee.

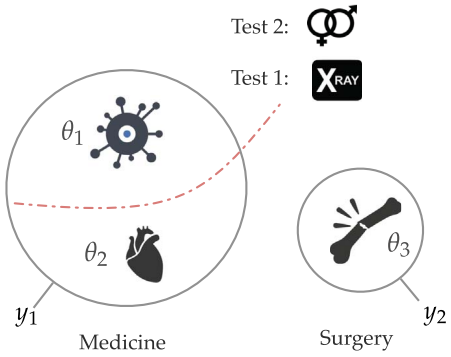
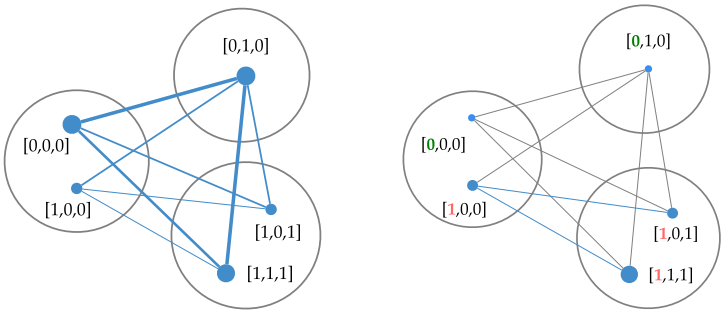


FIG 3. *An illustrative example for* EC^2*-Bayes and* ECED. *There are two tests: test* 1 *is very informative, as observing its outcome may immediately tell us which treatment is the best (e.g., if "X-ray result is negative", then we know* θ_3 *is false and hence the best treatment is* y_1*). Test* 2*, on the other hand, can be viewed as a "purely noisy" test, because knowing the gender doesn't change our belief of the root-causes. Hence, we want to design a criterion that encourages picking* x_1*.*

EC^2 with bayesian updates on edge weights. In the noisy setting, the test outcomes are not necessarily deterministic given a root-cause, i.e., $\forall\theta,\ \mathbb{P}[X_e \mid \theta] \in [0,1]$. Therefore, one can no longer "cut away" a root-cause θ by observing x_e, as long as $\mathbb{P}[X_e = x_e \mid \theta] > 0$. In such cases, a natural extension of the edge-cutting strategy will be – instead of cutting off edges – to *discount* the edge weights through Bayesian updates: After observing x_e, we can discount the weight of an edge (θ, θ'), by multiplying the probabilities of its incident root-causes with the likelihoods of the observation[4]: $w((\theta,\theta') \mid x_e) := \mathbb{P}[\theta]\,\mathbb{P}[\theta'] \cdot \mathbb{P}[x_e \mid \theta]\,\mathbb{P}[x_e \mid \theta'] = \mathbb{P}[\theta, x_e] \cdot \mathbb{P}[\theta', x_e]$. This gives us a greedy policy that, at every iteration, picks the test that has the maximal expected reduction in total edge weight. We call such policy EC^2-Bayes. Unfortunately, as we demonstrate later in §5, this seemingly promising update scheme is not ideal for solving our problem: it tends to pick tests that are very noisy, which do not help facilitate differentiation among different target values. Consider a simple example as illustrated in Fig. 3. There are three root-causes distributed as $\mathbb{P}[\theta_1] = 0.2, \mathbb{P}[\theta_2] = \mathbb{P}[\theta_3] = 0.4$, and two target values $r(\theta_1) = r(\theta_2) = y_1, r(\theta_3) = y_2$. We want to evaluate two tests: (1) a noiseless test X_1 with $\mathbb{P}[X_1 = 1 \mid \theta_1] = 1$ and $\mathbb{P}[X_1 = 1 \mid \theta_2] = \mathbb{P}[X_1 = 1 \mid \theta_3] = 0$; and (2) a purely noisy test X_2, i.e., $\forall\theta,\ \mathbb{P}[X_2 = 1 \mid \theta] = 0.5$. One can easily verify that by running EC^2-Bayes, one actually prefers X_2 (with expected reduction in edge weight 0.18, as opposed to 0.112 for X_1).

The ECED algorithm. The example above hints at an important principle of designing proper objective functions for this task: as the noise rate increases, one must take reasonable precautions when evaluating the informativeness of a

[4]Here we choose *not* to normalize the probabilities of θ, θ' to their posterior probabilities. Otherwise, we can end up having 0 gain in terms of edge weight reduction, even if we perform a very informative test.



(a) Initializing the ECED graph (b) Observing $X_1 = 1$.

FIG 4. *Illustration of the equivalence class edge discounting algorithm. Hypotheses are represented in dots. The size of a dot is proportional to its probabilities. Upon observing "inconsistent" outcomes, we discount the hypothesis accordingly and consequently discount its incident edges.*

test, such that the undesired contribution by noise is accounted for. Suppose we have performed test e and observed x_e. We call a root-cause θ to be "consistent" with observation x_e, if x_e is the most likely outcome of X_e given θ (i.e., $x_e \in \arg\max_x \mathbb{P}[X_e = x \mid \theta]$). Otherwise, we say θ is inconsistent. Now, instead of discounting the weight of all root-causes by the likelihoods $\mathbb{P}[X_e = x_e \mid \theta]$ (as EC^2-Bayes does), we choose to discount the root-causes by the *likelihood ratio*:

$$\lambda_{\theta,x_e} \triangleq \frac{\mathbb{P}[X_e = x_e \mid \theta]}{\max_{x'_e} \mathbb{P}[X_e = x'_e \mid \theta]}.$$

Intuitively, this is because we want to "penalize" a root-cause (and hence the weight of its incident edges), only if it is *inconsistent* with the observation (see Fig. 4). When x_e is consistent with root-cause θ, then $\lambda_{\theta,x_e} = 1$ and we do not discount θ; otherwise, if x_e is inconsistent with θ, we have $\lambda_{\theta,x_e} < 1$. When a test is not informative for root-cause θ, i.e. $\mathbb{P}[X_e \mid \theta]$ is uniform, then $\lambda_{\theta,x_e} = 1$, so that it neutralizes the effect of such test in terms of edge weight reduction.

Formally, given observations ψ_π, we define the (basic) value of observing x_e as the total amount of edge weight discounted:

$$\delta_{\text{BS}}(x_e \mid \psi_\pi) \triangleq \sum_{(\theta,\theta')\in E} \mathbb{P}[\theta, \psi_\pi]\,\mathbb{P}[\theta', \psi_\pi] \cdot (1 - \lambda_{\theta,x_e}\lambda_{\theta',x_e}).$$

Further, we call test e to be *non-informative*, if its outcome does not affect the distribution of Θ, i.e., $\forall\ \theta, \theta' \in \text{supp}(\Theta)$ and $x_e \in \mathcal{O}$, $\mathbb{P}[X_e = x_e \mid \theta] = \mathbb{P}[X_e = x_e \mid \theta']$. Obviously, performing a non-informative test does not reveal any useful information of Θ (and hence Y). Therefore, we should augment our basic value function δ_{BS}, such that the value of a non-informative test is 0. Following this principle, we define $\delta_{\text{OFFSET}}(x_e \mid \psi_\pi) \triangleq \sum_{(\theta,\theta')\in E} \mathbb{P}[\theta, \psi_\pi]\,\mathbb{P}[\theta', \psi_\pi] \cdot (1 - \max_\theta \lambda^2_{\theta,x_e})$, as the *offset* value for observing outcome x_e. It is easy to check that

Algorithm 1: The Equivalence Class Edge Discounting (ECED) algorithm

Input: $[\lambda_{\theta,x}]_{n\times m}$ (or Conditional Probabilities $\mathbb{P}[X \mid \Theta]$), Prior $\mathbb{P}[\Theta]$, Mapping $r : \mathrm{supp}(\Theta) \to \mathcal{Y}$;
begin
  $\psi_\pi \leftarrow \emptyset$;
  foreach $(\theta, \theta') \in E$ **do**
    $w_{\theta,\theta'} \leftarrow \mathbb{P}[\theta]\,\mathbb{P}[\theta']$;
  while $p_{\mathrm{ERR}}^{\mathrm{MAP}}(\psi_\pi) > \delta$ **do**
    $e^* \leftarrow \arg\max_e \mathbb{E}_{x_e}\Big[\sum_{(\theta,\theta')\in E} w_{\theta,\theta'} \cdot \big(\overbrace{1-\lambda_{\theta,x_e}\lambda_{\theta',x_e}}^{\text{weight discounted}} - \overbrace{(1-\max_{\theta''}\lambda^2_{\theta'',x_e})}^{\text{offset term}}\big)\Big]$;
    Observe x_{e^*}; $w_{\theta,\theta'} \leftarrow w_{\theta,\theta'} \cdot \mathbb{P}[x_{e^*} \mid \theta]\,\mathbb{P}[x_{e^*} \mid \theta']$;
    $\psi_\pi \leftarrow \psi_\pi \cup \{(e^*, x_{e^*})\}$;
  Output: $y^* = \arg\max_y \mathbb{P}[y \mid \psi_\pi]$.

if test e is non-informative, then it holds that $\delta_{\mathrm{BS}}(x_e \mid \psi_\pi) - \delta_{\mathrm{OFFSET}}(x_e \mid \psi_\pi) = 0$ for all $x_e \in \mathcal{O}$; otherwise $\delta_{\mathrm{BS}}(x_e \mid \psi_\pi) - \delta_{\mathrm{OFFSET}}(x_e \mid \psi_\pi) \geq 0$. This motivates us to use the following objective function:

$$\Delta_{\mathrm{ECED}}(X_e \mid \psi_\pi) \triangleq \mathbb{E}_{x_e}[\delta_{\mathrm{BS}}(x_e \mid \psi_\pi) - \delta_{\mathrm{OFFSET}}(x_e \mid \psi_\pi)], \tag{3.1}$$

as the expected amount of edge weight that is effectively reduced by performing test e. We call the algorithm that greedily maximizes Δ_{ECED} the *Equivalence Class Edge Discounting* (ECED) algorithm, and present the pseudocode in Algorithm 1.

Similar with EC^2, both the *computation complexity* (i.e., the running time) and the *query complexity* (i.e., number of tests needed) of ECED depend on the number of root-causes. Let $\epsilon_{\theta,e} \triangleq 1 - \max_x \mathbb{P}[X_e = x \mid \theta]$ be the noise rate for test e. As our main theoretical result, we show that under the basic setting where test outcomes are *binary*, and the test noise is *independent* of the underlying root-causes (i.e., $\forall\theta \in \mathrm{supp}(\Theta)$, $\epsilon_{\theta,e} = \epsilon_e$), ECED is competitive with the optimal policy which achieves a lower error probability for Problem (2.1):

Theorem 1. *Let $\delta \in (0,1)$ be the target error probability which is achievable. To achieve expected error probability less than δ, it suffices to run* ECED

$$O\left(\frac{k}{c_\varepsilon}\left(\log\frac{kn}{\delta}\log\frac{n}{\delta}\right)^2\right)$$

steps. Here, $n \triangleq |\mathrm{supp}(\Theta)|$ is the number of root-causes, $c_\varepsilon \triangleq \min_{e\in\mathcal{V}}(1-2\epsilon_e)^2$ characterizes the severity of noise, and $k \triangleq \min\{m, \mathrm{cost}(OPT(\delta_{\mathsf{opt}}))\}$, where $\mathrm{cost}(OPT(\delta_{\mathsf{opt}}))$ denotes the worst-case cost of the optimal policy that achieves expected error probability $\delta_{\mathsf{opt}} \triangleq O\left(\frac{\delta}{(\log n\cdot\log(1/\delta))^2}\right)$.

Note that a pessimistic upper bound for k is the total number of tests m, and hence the cost of ECED is at most $O\left((\log(mn/\delta)\log(n/\delta))^2/c_\varepsilon\right)$ times the

worst-case cost of the optimal algorithm, which achieves a lower error probability $O\left(\delta/(\log n \cdot \log(1/\delta))^2\right)$. Further, as one can observe, the upper bound on the cost of ECED degrades as we increase the maximal noise rate of the tests. When $c_\varepsilon = 1$, we have $\epsilon_e = 0$ for all test e, and ECED reduces to the EC^2 algorithm. Theorem 1 implies that running EC^2 for $O\left(k\left(\log \frac{kn}{\delta} \log \frac{n}{\delta}\right)^2\right)$ in the noise-free setting is sufficient to achieve $p_{\mathrm{ERR}} \leq \delta$. Finally, notice that by construction ECED never selects any non-informative test. Therefore, we can always remove purely noisy tests (i.e., $\{e : \forall\theta,\ \mathbb{P}[X_e = 1 \mid \theta] = \mathbb{P}[X_e = 0 \mid \theta] = 1/2\}$), so that $c_\varepsilon > 0$, and the upper bound in Theorem 1 becomes non-trivial.

4. Theoretical analysis

Information-theoretic auxiliary function. We now present the main idea behind the proof of Theorem 1. In general, an effective way to relate the performance (measured in terms of the gain in the target objective function) of the greedy policy to the optimal policy is by showing that, the *one-step* gain of the greedy policy always makes effective progress towards approaching the cumulative gain of OPT *over k steps*. One powerful tool facilitating this is the *adaptive submodularity* theory, which imposes a lower bound on the one-step greedy gain against the optimal policy, given that the objective function in consideration exhibits a natural diminishing returns condition. Unfortunately, in our context, the target function to optimize, i.e., the expected error probability of a policy, does not satisfy adaptive submodularity. Furthermore, it is nontrivial to understand how one can directly relate the two objectives: the ECED objective of (3.1), which we utilize for selecting informative tests, and the gain in the reduction of error probability, which we use for evaluating a policy.

We circumvent such problems by introducing auxiliary functions, as a proxy to connect the ECED objective Δ_{ECED} with the expected reduction in error probability p_{ERR}. Ideally, we aim to find some auxiliary objective f_{AUX}, such that the tests with the maximal Δ_{ECED} also have a high gain in f_{AUX}; meanwhile, f_{AUX} should also be comparable with the error probability p_{ERR}, such that minimizing f_{AUX} itself is sufficient for achieving low error probability.

We consider the function $f_{\mathrm{AUX}} : 2^{\mathcal{V}\times\mathcal{O}} \to \mathbb{R}_{\geq 0}$, defined as

$$f_{\mathrm{AUX}}(\psi) = \sum_{(\theta,\theta')\in E} \mathbb{P}[\theta \mid \psi]\,\mathbb{P}[\theta' \mid \psi] \cdot \log \frac{1}{\mathbb{P}[\theta \mid \psi]\,\mathbb{P}[\theta' \mid \psi]} + c \sum_{y\in\mathcal{Y}} \mathbb{H}_2\left(\mathbb{P}[y \mid \psi]\right). \tag{4.1}$$

Here $\mathbb{H}_2(x) = -x\log x - (1-x)\log(1-x)$ denotes the binary entropy function, and c is a constant that will be made concrete shortly (in Lemma 3). Interestingly, we show that function f_{AUX} is intrinsically linked to the error probability:

Lemma 2. *We consider the auxiliary function defined in Equation (4.1). Let* $n \triangleq |\mathrm{supp}(\Theta)|$ *be the number of root-causes, and* $p_{\mathrm{ERR}}^{MAP}(\psi)$ *be the error proba-*

bility given partial realization ψ. Then

$$2c \cdot p_{\mathrm{ERR}}^{MAP}(\psi) \leq f_{\mathrm{AUX}}(\psi) \leq (3c+4) \cdot \left(\mathbb{H}_2\left(p_{\mathrm{ERR}}^{MAP}(\psi)\right) + p_{\mathrm{ERR}}^{MAP}(\psi) \log n\right).$$

Therefore, if we can show that by running ECED we can effectively reduce f_{AUX}, then by Lemma 2, we can conclude that ECED also makes significant progress in reducing the error probability $p_{\mathrm{ERR}}^{\mathrm{MAP}}$.

Bounding the gain w.r.t. the auxiliary function. It remains to understand how ECED interacts with f_{AUX}. For any test e, we define $\Delta_{\mathrm{AUX}}(X_e \mid \psi) \triangleq \mathbb{E}_{x_e}[f_{\mathrm{AUX}}(\psi \cup \{e, x_e\}) - f_{\mathrm{AUX}}(\psi) \mid \psi]$ to be the expected gain of test e in f_{AUX}. Let $\Delta_{\mathrm{EC}^2,\psi}(X_e)$ denote the gain of test e in the EC^2 objective, assuming that the edge weights are configured according to the *posterior distribution* $\mathbb{P}[\Theta \mid \psi]$. Similarly, let $\Delta_{\mathrm{ECED},\psi}(X_e)$ denote the ECED gain, if the edge weights are configured according to $\mathbb{P}[\Theta \mid \psi]$. We prove the following result:

Lemma 3. *Let $n = |\operatorname{supp}(\Theta)|$, $t = |\mathcal{Y}|$, and ϵ be the noise rate associated with test $e \in \mathcal{V}$. Fix $\eta \in (0,1)$. We consider f_{AUX} as defined in Equation (4.1), with $c = 8\left(\log(2n^2/\eta)\right)^2$. It holds that*

$$\Delta_{\mathrm{AUX}}(X_e \mid \psi) + c_{\eta,\epsilon} \geq \Delta_{\mathrm{ECED},\psi}(X_e) \cdot (1-\epsilon)^2/16 = c_\epsilon \Delta_{\mathrm{EC}^2,\psi}(X_e),$$

where $c_{\eta,\epsilon} = 2t(1-2\epsilon)^2\eta$, and $c_\epsilon \triangleq (1-2\epsilon)^2/16$.

Lemma 3 indicates that the test being selected by ECED can effectively reduce f_{AUX}.

Lifting the adaptive submodularity framework. Recall that our general strategy is to bound the one step gain in f_{AUX} against the gain of an optimal policy. In order to do so, we need to show that our surrogate exhibits, to some extent, the diminishing returns property. By Lemma 3 we can relate $\Delta_{\mathrm{AUX}}(X_e \mid \psi_\pi)$, i.e., the gain in f_{AUX} under the *noisy* setting, to $\Delta_{\mathrm{EC}^2,\psi}(X_e)$, i.e., the expected weight of edges *cut* by the EC^2 algorithm. Since f_{EC^2} is adaptive submodular, this allows us to lift the adaptive submodularity framework into the analysis. As a result, we can now relate the 1-step gain w.r.t. f_{AUX} of a test selected by ECED, to the cumulative gain w.r.t. f_{EC^2} of an optimal policy in the noise-free setting. Further, observe that the EC^2 objective at ψ satisfies:

$$f_{\mathrm{EC}^2,\psi} := \sum_y \mathbb{P}[y \mid \psi]\left(1 - \mathbb{P}[y \mid \psi]\right) \overset{(a)}{\geq} 1 - \max_y \mathbb{P}[y \mid \psi] = p_{\mathrm{ERR}}^{\mathrm{MAP}}(\psi). \tag{4.2}$$

Hereby, step (a) is due to the fact that the error probability of a MAP estimator always lower bounds that of a stochastic estimator (which is drawn randomly according to the posterior distribution of Y). Suppose we want to compare ECED against an optimal policy OPT. By adaptive submodularity, we can relate the 1-step gain of ECED in $f_{\mathrm{EC}^2,\psi}$ to the cummulative gain of OPT. Combining Equation (4.2) with Lemma 2 and Lemma 3, we can bound the 1-step gain in f_{AUX} of ECED against the k-step gain of OPT, and consequently bound the cost of ECED against OPT for Problem 2.1. We defer a more detailed proof outline and the full proof to Appendix B.

5. Experimental results

We now demonstrate the performance of ECED on three real-world problem instances: a Bayesian experimental design task intended to distinguish among economic theories of how people make risky decisions, an active preference learning task via pairwise comparisons, and a third case study on pool-based active learning for classification.

Baselines. The first baseline we consider is EC^2-Bayes, which uses the Bayes' rule to update the edge weights when computing the gain of a test (as described in §3). Note that after observing the outcome of a test, both ECED and EC^2-Bayes update the posteriors on Θ and Y according to the Bayes' rule; the only difference is that they use different strategies when *selecting* a test. We also compare with two commonly used sequential information gathering policies: Information Gain (IG) and Uncertainty Sampling (US), which consider picking tests that greedily maximizing the reduction of entropy over the target variable Y, and root-causes Θ respectively. Last, we consider myopic optimization of the decision-theoretic value of information (VOI) [19]. In our problems, the VOI policy greedily picks the test maximizing the expected reduction in prediction error in Y.

5.1. Preference elicitation in behavioral economics

We first conduct experiments on a Bayesian experimental design task, which intends to distinguish among economic theories of how people make risky decisions. Several theories have been proposed in behavioral economics to explain how people make decisions under risk and uncertainty. We test ECED on six theories of subjective valuation of risky choices [33, 32, 30], namely (1) *expected utility with constant relative risk aversion*, (2) *expected value*, (3) *prospect theory*, (4) *cumulative prospect theory*, (5) *weighted moments*, and (6) *weighted standardized moments*. Choices are between risky lotteries, i.e., known distribution over payoffs (e.g., the monetary value gained or lost). A test $e \triangleq (L_1, L_2)$ is a pair of lotteries, and root-causes Θ correspond to parametrized theories that predict, for a given test, which lottery is preferable. The goal, is to adaptively select a sequence of tests to present to a human subject to distinguish which of the six theories best explains the subject's responses. We employ the same set of parameters used in [26] to generate tests and root-causes. In particular, we have generated ∼16K tests. Given root-cause θ and test $e = (L_1, L_2)$, one can compute the values of L_1 and L_2, denoted by v_1 and v_2. The noise of a test is characterized by the Bradley-Terry-Luce (BTL) preference model[5] [3], where the probability that root-cause θ favors L_1 is defined as $\mathbb{P}[X_e = 1 \mid \theta] = \frac{1}{1+\exp(-\lambda\cdot(v_1 - v_2))}$.

[5]The BTL model has been widely used for pairwise data, e.g., [24, 29], etc. Intuitively, the user is more prone to error if the utilities of a pair are close. I.e., for preference elicitation, if a pair of lotteries (L_1, L_2) is almost of equal value to the user, then her feedback on whether she favors L_1 over L_2 is very noisy.

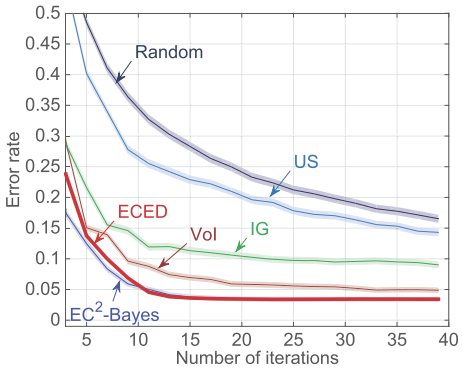
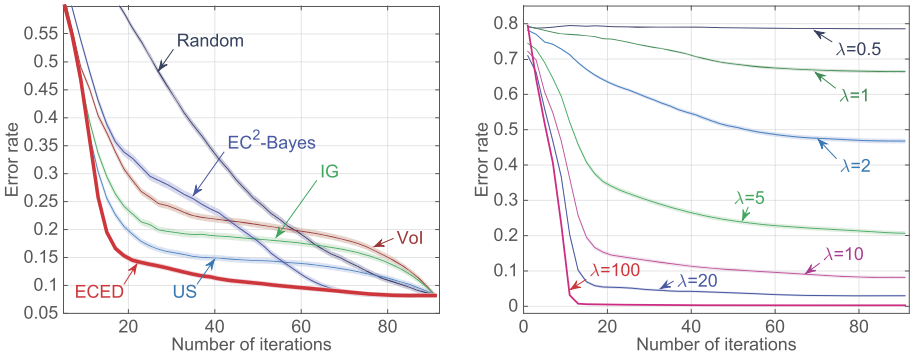


FIG 5. *Experimental results:* ECED *outperforms most baselines on both data sets.*

Results. To evaluate ECED, we do not specify a target error probability δ as input. Instead, we set a budget on the number of iterations allowed, and plot the error probability as a function of the number of iterations. Fig. 5 demonstrates the performance of ECED. The average error probability has been computed across 1000 random trials for all methods. We observe that ECED and EC^2-Bayes have similar behavior on this data set; however, the performance of the US algorithm is much worse. This can be explained by the nature of the data set: it has more concentrated distribution over Θ, but not Y. Therefore, since tests only provide indirect information about Y through Θ, what the uncertainty sampling scheme tries to optimize is actually Θ, hence it performs quite poorly.

5.2. Active preference learning via pairwise comparisons

The second application considers a comparison-based movie recommendation system, which learns a user's movie preference (e.g., the favorable genre) by sequentially showing her pairs of candidate movies, and letting her choose which one she prefers. We use the *MovieLens 100k* dataset [18] which consists of a matrix of 1 to 5 ratings of 1682 movies from 943 users, and adopt the experimental setup proposed in [8]. In particular, we extract movie features by computing a low-rank approximation of the user/rating matrix of the *MovieLens 100k* dataset through singular value decomposition (SVD). We then simulate the target "categories" Y that a user may be interested by partitioning the set of movies into t (non-overlapping) clusters in the Euclidean space. A root-cause Θ corresponds to user's favorite movie, and tests e's are given in the form of movie pairs, i.e., $e \triangleq (a, b)$, where a and b are embeddings of the two movies in Euclidean space. Suppose user's movie is represented by θ, then test e is realized as 1 if a is closer to y than b, and 0 otherwise. Similarly with the previous application, we model the noise with the BTL model, i.e., $\mathbb{P}[X_e = 1 \mid \theta] = \frac{1}{1+\exp(-\lambda\cdot(d(m_a,\theta)-d(m_b,\theta)))}$.



(a) *MovieLens* (b) *MovieLens* - Varying noise

FIG 6. *Experimental results on active preference learning*

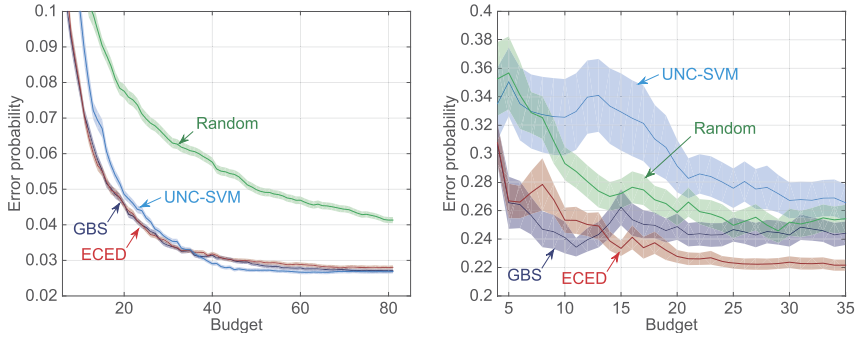
where $d(\cdot,\cdot)$ is the distance function, and λ controls the level of noise in the system.

Results. Fig. 6(a) shows the performance of ECED compared to other baseline methods when we fix $|\mathcal{Y}| = 20$ and $\lambda = 10$. We compute the average error probability across 1000 random trials for all methods. We can see that ECED consistently outperforms all other baselines. Interestingly, EC^2-Bayes performs poorly on this data set. This could be because the noise level is still high, misguiding the two heuristics to select noisy, uninformative tests. Fig. 6(b) shows the performance of ECED as we vary λ. When $\lambda = 100$, the tests become close to deterministic given a root-cause, and ECED can achieve 0 error with ~ 12 tests. As we increase the noise rate (i.e., decrease λ), it takes ECED many more queries for the prediction error to converge. This is because with high noise rate, ECED discounts the root-causes more uniformly, and therefore those tests are hardly informative about Y. It comes at the cost of performing more tests, and hence low convergence rate.

5.3. Pool-based active learning

To demonstrate the empirical performance of ECED, we further conduct experiments on two pool-based binary active classification tasks. In the active learning application, we can sequentially query from a pool of data points, and the goal is to learn a binary classifier, which achieves some small prediction error on the unseen data points from the pool, with the smallest possible number of queries.

Active learning: Targets and root-causes. To discretize the hypotheses space, we use a noisy version of hit-and-run sampler as suggested in Chen and Krause [6]. Each hypothesis can be represented by a binary vector indicating the



(a) *WDBC* (b) *Fourclass*

FIG 7. *Pool-based Active Learning for Classification*

outcomes of all data points in the training set. Then, we construct an epsilon-net on the set of hypotheses (based on the Hamming distance between hypotheses). We obtain the equivalence classes for ECED, by assigning each hypothesis to its closest center of epsilon-ball, measured by their Hamming distances. Note that the Hamming distance between two hypotheses reflects the difference of prediction error. Consider epsilon-net of fixed radius ε. By construction, hypotheses that lie in the some equivalence classes are at most 2ε away from each other; therefore the hypotheses which are within the epsilon-ball of the optimal hypotheses are considered to be near-optimal. Using the terminology in this paper, hypotheses correspond to root-causes, and the groups of hypothesis correspond to the target variable of interest. Running ECED, ideally, will help us locate a near-optimal epsilon-ball as quickly as possible.

Baselines. We compare ECED with the popular uncertainty sampling heuristic (UNC-SVM), which sequentially queries the data points which are the closest to the decision boundary of a SVM classifier. We also compare with the GBS algorithm, which sequentially queries the data points that maximally reduces the volume of the version space.

Results. We evaluate ECED and the baseline algorithms on the *UCI WDBC* dataset (569 instances, 32-d) and *Fourclass* dataset (862 instances, 2-d). For ECED and GBS, we sample a fixed number of 1000 hypotheses in each random trial. For both instances we assume a constant error rate $\epsilon = 0.02$ for all tests. Fig. 7(a) and Fig. 7(b) demonstrate that ECED is competitive with the baselines. Such results suggests that grouping of hypotheses could be beneficial when learning under noisy data.

6. Related work

Active learning in statistical learning theory. In most of the theoretical active learning literature (e.g., Dasgupta [10], Hanneke [15, 16], Balcan

and Urner [1]), active learning algorithms are mainly considered in terms of their *statistical complexity*, which is defined as the worst-case cost (i.e., number of labels) required to learn a classifier with classification error below certain threshold. Bounds on statistical complexity have been characterized in terms of the structure of the hypothesis class, as well as additional distribution-dependent complexity measures (e.g., splitting index [10], disagreement coefficient [15, 34], etc); In comparison, in this paper we study the problem of exact identification of the target random variable, and focus on the optimality: we seek computationally-efficient approaches that are *provably competitive* with the optimal policy. Therefore, we do not seek to bound how the optimal policy behaves, and hence we make no assumptions on the hypothesis class (e.g., we don't restrict $\mathcal{Y}$ or $\text{supp}(\Theta)$ to be a set of linear classifiers).

Persistent noise vs. non-persistent noise. If tests can be repeated with i.i.d. outcomes, the noisy problem can then be effectively reduced to the noise-free setting [20, 22, 25]. While the modeling of non-persistent noise may be appropriate in some settings (e.g., if the noise is due to measurement error), it is often important to consider the setting of *persistent noise*: In many applications, repeating tests are impossible or produces identical outcomes. For example, it could be unrealistic to replicate a medical test for practical clinical treatment. Despite some recent development in dealing with persistent noise in simple graphical models [7] and strict noise assumptions [14], more general settings, which we focus on in this paper, are much less understood.

7. Conclusion

We have introduced ECED, which strictly generalizes the EC^2 algorithm, for solving practical Bayesian active learning and experimental design problems with correlated and noisy tests. By introducing an analysis framework that draws upon adaptive submodularity and information theory, we have proved that ECED enjoys strong theoretical guarantees. We have demonstrated the compelling performance of ECED on two (noisy) problem instances, including an active preference learning task via pairwise comparisons, and a Bayesian experimental design task for preference elicitation in behavioral economics. We believe that our work makes an important step towards understanding the theoretical aspects of complex, sequential information gathering problems, and provides useful insight on how to develop practical algorithms to address noise.

Appendix A: Table of notations defined in the main paper

We summarize the notations used in the main paper in Table 1.

TABLE 1
A reference table of notations used in the main paper

Y, y	random variable encoding the value of the target variable and its value
$\mathcal{Y}$	domain of the target variable
Θ, θ	random variable encoding the root-cause, and its realization
$\text{supp}(\Theta)$	the ground set / domain of root-causes
r	$\Theta \to Y$, a function that maps a root-cause to a target value
$\mathcal{V}$, e	the ground set of tests, and the index of a test
m	$\lvert\mathcal{V}\rvert$, number of tests
X_e, x_e	random variable encoding the test outcome and its realization
t	$\lvert\mathcal{Y}\rvert$, number of possible target values
n	$\lvert\text{supp}(\Theta)\rvert$, number of root-causes
π	policy, i.e., a (partial) mapping from observation vectors to tests
Ψ, ψ_π	random variable encoding a partial realization and its value.
δ	tolerance of the (expected) error probability
$p_{\text{ERR}}^{\text{MAP}}(\psi)$	error probability (of a MAP decoder), having observed ψ
$p_{\text{ERR}}(\pi)$	$\mathbb{E}_{\psi_\pi}\left[p_{\text{ERR}}^{\text{MAP}}(\psi_\pi)\right]$, expected error probability by running policy π
OPT	optimal policy for Problem (2.1)
G	$G = (\text{supp}(\Theta), E)$, the graph constructed for the EC^2 algorithm
$w(\{\theta, \theta'\})$	weight of edge $\{\theta, \theta'\} \in E$ in the EC^2 graph G
f_{EC^2}	EC^2 objective, with $f_{\text{EC}^2}(\emptyset) := \sum_{\theta,\theta' \in E} \mathbb{P}[\theta]\,\mathbb{P}[\theta']$.
$f_{\text{EC}^2,\psi}$	EC^2 objective, with $f_{\text{EC}^2,\psi}(\emptyset) := \sum_{\theta,\theta' \in E} \mathbb{P}[\theta \mid \psi]\,\mathbb{P}[\theta' \mid \psi]$.
$\lambda_{\theta,e}$	discount coefficient of θ, used by ECED when computing Δ_{ECED}.
$\epsilon_{\theta,e}$	$1 - \arg\max_e \mathbb{P}[X_e = x_e]$, the noise rate for a test e
$\delta_{\text{BS}}(x_e \mid \psi)$	the "basic" component in the ECED gain of x_e, having observed ψ
$\delta_{\text{OFFSET}}(x_e \mid \psi)$	the "offset" component in the ECED gain of x_e, having observed ψ
$\Delta_{\text{ECED}}(X_e \mid \psi)$	the ECED gain which is myopically optimized.
$\Delta_{\text{ECED},\psi}(X_e)$	suppose we have observed ψ, and re-initialize the EC^2 graph so that the total edge weight is $f_{\text{EC}^2,\psi}(\emptyset)$. Then, $\Delta_{\text{EC}^2,\psi}(X_e)$ is the expected reduction in edge weight, by performing test e and *discounting* edges' weight according to ECED. It is the re-normalized version of $\Delta_{\text{ECED}}(x_e \mid \psi)$, i.e., $\Delta_{\text{ECED},\psi}(X_e) = \Delta_{\text{ECED}}(x_e \mid \psi)/\mathbb{P}[\psi]^2$.
$\Delta_{\text{EC}^2,\psi}(X_e)$	the expected gain in $f_{\text{EC}^2,\psi}$ by performing test e, and *cutting* edges weight according to EC^2. It can be interpreted as $\Delta_{\text{ECED},\psi}(X_e)$, as if the test's outcome is noise-free, i.e., $\forall\theta,\ \epsilon_{\theta,e} = 0$.
f_{AUX}	the auxiliary function defined in Equation (4.1)
η	parameter of f_{AUX} (see Eq (4.1), Lemma 3). It is only used for analysis.
c	$8\left(\log(n^2/\eta)\right)^2$, parameter of f_{AUX}. It is only used for the analysis.
$\Delta_{\text{AUX}}(X_e \mid \psi)$	the expected gain in f_{AUX} by running e given partial realization ψ
$c_{\eta,\epsilon}, c_\epsilon$	constants required by Lemma 3
λ	parameter controlling the error rate of tests (see §5)

Appendix B: The analysis framework

In this section, we provide the proofs of our theoretical results in full detail. Recall that for the theoretical analysis, we study the basic setting where test outcomes are *binary*, and the test noise is *independent* of the underlying root-causes (i.e., given a test e, the noise rate on the outcome of test e is only a function of e, but not a function of θ).

B.1. Proof of Theorem 1 outline: Introducing auxiliary functions

The general idea behind our analysis, is to show that by running ECED, the one-step gain in learning the value of the target variable is significant, compared with the cumulative gain of an optimal policy over k steps (see Fig. 8).

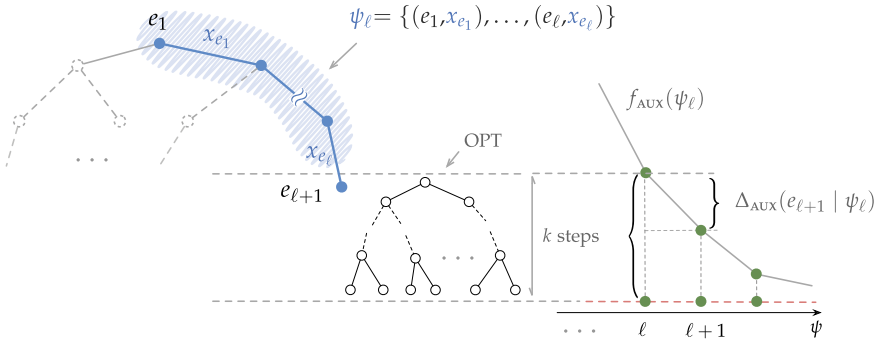


FIG 8. *On the left, we demonstrate a sequential policy in the form of its decision tree representation. Nodes represent tests selected by the policy, and edges represent outcomes of tests. At step ℓ, a policy maps partial realization $\psi_\ell = \{(e_1, x_{e_1}), \ldots, (e_\ell, x_{e_\ell})\}$ to the next test $e_{\ell+1}$ to be performed. In the middle, we demonstrate the tests selected by an optimal policy* OPT *of length k. On the right, we illustrate the change in the auxiliary function as* ECED *selects more tests. Running* OPT *at any step of execution of* ECED *will make f_{AUX} below some threshold (represented by the red dotted line). The key idea behind our proof, is to show that the greedy policy* ECED*, at each step, is making effective progress in reducing the expected prediction error (in the long run), compared with* OPT*.*

In Appendix §C, we show that if tests are greedily selected to optimize the (reduction in) expected prediction error, we may end up failing to pick some tests, which have negligible immediate gain in terms of error reduction, but are very informative in the long run. ECED bypasses such an issue by selecting tests that maximally distinguish root-causes with different target values. In order to analyze ECED, we need to find an auxiliary function that properly tracks the "progress" of the ECED algorithm; meanwhile, this auxiliary function should allow us to connect the heuristic by which we select tests (i.e., Δ_{ECED}), with the target objective of interest (i.e., the expected prediction error p_{ERR}).

We consider the auxiliary function defined in Equation (4.1). For brevity, we suppress the dependence of ψ where it is unambiguous. Further, we use p_θ, $p_{\theta'}$, and p_y as shorthand notations for $\mathbb{P}[\theta \mid \psi]$, $\mathbb{P}[\theta' \mid \psi]$ and $\mathbb{P}[y \mid \psi]$. Equation (4.1) can be simplified as

$$f_{\text{AUX}} = \sum_{(\theta,\theta') \in E} p_\theta p_{\theta'} \log \frac{1}{p_\theta p_{\theta'}} + c \sum_{y \in \mathcal{Y}} \mathbb{H}_2(p_y) \tag{B.1}$$

We illustrate the outline of our proofs in Fig. 9. Our goal is to bound the cost of ECED against the cost of OPT (Theorem 1; proof provided in Appendix §B.5). As we have explained earlier, our strategy is to relate the one-step gain

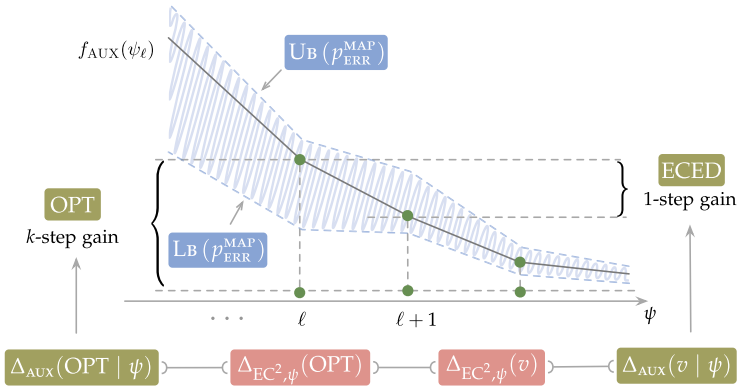


FIG 9. *The proof outline.*

of ECED 1-step: $\Delta_{\text{AUX}}(e_{\ell+1} \mid \psi_\ell)$ with the gain of OPT in k-steps $\Delta_{\text{AUX}}(\text{OPT} \mid \psi)$ (Appendix §B.4.2, Lemma 8). To achieve that, we divide our proof into three parts:

Part 1 We show that the auxiliary function f_{AUX} is closely related with the target objective function p_{ERR}. More specifically, we provide both an upper bound UB $(p^{\text{MAP}}_{\text{ERR}})$ and a lower bound LB $(p^{\text{MAP}}_{\text{ERR}})$ of f_{AUX} in Lemma 2, and give the detailed proofs in Appendix §B.2.

Part 2 To analyze the one-step gain of ECED, we introduce another intermediate auxiliary function: For a test $x_{\ell+1}$ chosen by ECED, we relate its one-step gain in the auxiliary function $\Delta_{\text{AUX}}(v \mid \psi)$, to its one-step gain in the EC^2 objective $\Delta_{\text{EC}^2,\psi}(v)$ (Lemma 3, detailed proof provided in Appendix §B.3). The reason why we introduce this step is that the EC^2 objective is adaptive submodular, by which we can relate the 1-step gain of a greedy policy $\Delta_{\text{EC}^2,\psi}(v)$ to an optimal policy $\Delta_{\text{EC}^2,\psi}(\text{OPT})$.

Part 3 To close the loop, it remains to connect the gain of an optimal policy OPT in the EC^2 objective function $\Delta_{\text{EC}^2,\psi}(\text{OPT})$, with the gain of OPT in the auxiliary function $\Delta_{\text{AUX}}(\text{OPT} \mid \psi)$. We establish such connection in Lemma 8, and present its proof in Appendix §B.4.1.

To make the proof more accessible, we insert the annotated color blocks from Fig. 9 (i.e., UB $(p^{\text{MAP}}_{\text{ERR}})$, LB $(p^{\text{MAP}}_{\text{ERR}})$, $\Delta_{\text{AUX}}(v \mid \psi)$, $\Delta_{\text{EC}^2,\psi}(v)$, $\Delta_{\text{EC}^2,\psi}(\text{OPT})$, $\Delta_{\text{AUX}}(\text{OPT} \mid \psi)$, etc), into the subsequent subsections in Appendix §B, so that readers can easily relate different parts of this section to the proof outline. Note that we only use these annotated color blocks for positioning the proofs, and hence readers can ignore the notations, as it may slightly differ from the ones used in the proof.

B.2. Proof of Theorem 1 part 1: Proof of Lemma 2

In this subsection, we provide the proof of Lemma 2, which relates f_{AUX} to p_{ERR}.

Define $p_{\mathrm{E}}(\psi) \triangleq \sum_{y\in\mathcal{Y}} \mathbb{P}[y \mid \psi](1-\mathbb{P}[y \mid \psi])$ as the prediction error of a *stochastic estimator* upon observing ψ, i.e., the probability of mispredicting y if we make a random draw from $\mathbb{P}[Y \mid \psi]$. We show in Lemma 4 that $p_{\mathrm{ERR}}^{\mathrm{MAP}}(\psi)$ is within a constant factor of $p_{\mathrm{E}}(\psi)$:

Lemma 4. *Fix ψ, it holds that $p_{\mathrm{ERR}}^{\mathrm{MAP}}(\psi) \le p_{\mathrm{E}}(\psi) \le 2p_{\mathrm{ERR}}^{\mathrm{MAP}}(\psi)$.*

Proof of Lemma 4. We can always lower bound p_{E} by $p_{\mathrm{ERR}}^{\mathrm{MAP}}$, since by definition, $p_{\mathrm{ERR}}^{\mathrm{MAP}}(\psi) = 1-\max_y \mathbb{P}[y \mid \psi] = \sum_{y\in\mathcal{Y}} \mathbb{P}[y \mid \psi]\cdot(1-\max_y \mathbb{P}[y \mid \psi]) \le \sum_{y\in\mathcal{Y}} \mathbb{P}[y \mid \psi](1-\mathbb{P}[y \mid \psi]) = p_{\mathrm{E}}(\psi)$.

To prove the second part, we write $p_{y_i} = \mathbb{P}[Y = y_i \mid \psi]$ for all $y_i \in \mathcal{Y}$. W.l.o.g., we assume $p_{y_1} \ge p_{y_2} \ge \cdots \ge p_{y_t}$. Then $p_{\mathrm{ERR}}^{\mathrm{MAP}} = 1-p_{y_1}$. We further have

$$\begin{aligned}
2p_{\mathrm{ERR}}^{\mathrm{MAP}} &= 2(1-p_{y_1}) = 2(\sum_{i=2}^{t} p_{y_i}) = 2(\sum_{i=1}^{t} p_{y_i})(\sum_{i=2}^{t} p_{y_i}) \\
&= 2(p_{y_1} + \sum_{i=2}^{t} p_{y_i})(\sum_{i=2}^{t} p_{y_i}) \\
&\ge 2p_{y_1}(\sum_{i=2}^{t} p_{y_i}) + (\sum_{i=2}^{t} p_{y_i})^2 \\
&\ge \sum_{i\ne j}^{t} p_{y_i}p_{y_j} = \sum_{i} p_{y_i}(1-p_{y_i}) = p_{\mathrm{E}}
\end{aligned}$$

□

Now, we provide lower and upper bounds of the second term in the RHS of Equation (B.1):

Lemma 5. $2p_{\mathrm{ERR}}^{\mathrm{MAP}} \le \sum_{y\in\mathcal{Y}} \mathbb{H}_2(p_y) \le 3(\mathbb{H}_2(p_{\mathrm{ERR}}^{\mathrm{MAP}}) + p_{\mathrm{ERR}}^{\mathrm{MAP}} \log n)$.

Proof of Lemma 5. We first prove the inequality on the left. Expanding the middle term involving the binary entropy of p_y, we get

$$\begin{aligned}
\sum_{y\in\mathcal{Y}} \mathbb{H}_2(p_y) &= \sum_{y\in\mathcal{Y}} \left(p_y \log\frac{1}{p_y} + (1-p_y)\log\frac{1}{1-p_y} \right) \\
&\overset{\text{(a)}}{\ge} \frac{2}{\ln 2} \sum_{y\in\mathcal{Y}} p_y(1-p_y) \\
&\ge 2p_{\mathrm{E}} \overset{\text{Lemma 4}}{\ge} 2p_{\mathrm{ERR}}^{\mathrm{MAP}}
\end{aligned}$$

Here, step (a) is by inequality $\ln x \ge 1-1/x$ for $x \ge 0$.

To prove the second part, we first show in the following that

$$\sum_{y}(1-p_y)\log\frac{1}{1-p_y} \le 2\sum_{y} p_y \log\frac{1}{p_y}.$$

W.l.o.g., we assume that the probabilities p_y's are in decreasing order, i.e., $p_{y_1} \ge p_{y_2} \ge \cdots \ge p_{y_t}$. Observe that if $p_y \in [0, 1/2]$, then $(1-p_y)\log\frac{1}{1-p_y} \le p_y \log\frac{1}{p_y}$. Consider the following two cases:

1. $p_{y_1} \le 1/2$. In this case, we have $\sum_y (1-p_y)\log\frac{1}{1-p_y} \le \sum_y p_y \log\frac{1}{p_y}$.
2. $p_{y_1} > 1/2$. Since $\sum_{i>1} p_{y_i} = 1 - p_{y_1}$, we have

$$\begin{aligned}
\sum_i (1-p_{y_i})\log\frac{1}{1-p_{y_i}} &= (1-p_{y_1})\log\frac{1}{1-p_{y_1}} + \sum_{i>1}(1-p_{y_i})\log\frac{1}{1-p_{y_i}} \\
&= \sum_{i>1} p_{y_i}\log\frac{1}{\sum_{i>1} p_{y_i}} + \sum_{i>1}(1-p_{y_i})\log\frac{1}{1-p_{y_i}} \\
&\le \sum_{i>1} p_{y_i}\log\frac{1}{p_{y_i}} + \sum_{i>1}(1-p_{y_i})\log\frac{1}{1-p_{y_i}} \\
&\le \sum_{i>1} p_{y_i}\log\frac{1}{p_{y_i}} + \sum_{i>1} p_{y_i}\log\frac{1}{p_{y_i}} \\
&\le 2\sum_{i>0} p_{y_i}\log\frac{1}{p_{y_i}}
\end{aligned}$$

Therefore,

$$\sum_{y\in\mathcal{Y}} \mathbb{H}_2(p_y) \le 3\sum_{i>0} p_{y_i}\log\frac{1}{p_{y_i}} = 3\mathbb{H}(Y). \tag{B.2}$$

Furthermore, by Fano's inequality (in the absence of conditioning), we know that $\mathbb{H}(Y) \le \mathbb{H}_2\left(p_{\text{ERR}}^{\text{MAP}}\right) + p_{\text{ERR}}^{\text{MAP}}\log(|\mathcal{Y}|-1)$. Combining with Equation (B.2) we get

$$\sum_y \mathbb{H}_2(p_y) \le 3\mathbb{H}(Y) \le 3\left(\mathbb{H}_2\left(p_{\text{ERR}}^{\text{MAP}}\right) + \log(|\mathcal{Y}|-1)\right) \overset{(b)}{\le} 3\left(\mathbb{H}_2\left(p_{\text{ERR}}^{\text{MAP}}\right) + \log(n)\right)$$

where in (b) we use the fact that $t = |\mathcal{Y}| \le |\operatorname{supp}(\Theta)| = n$, since $Y = r(\Theta)$ is a function of Θ. Hence it completes the proof. □

Next, we bound the first term on the RHS of Equation (B.1), i.e.,

$$\sum_{\{\theta,\theta'\}\in E} p_\theta p_{\theta'} \log\frac{1}{p_\theta p_{\theta'}},$$

against $p_{\text{ERR}}^{\text{MAP}}$:

Lemma 6. $\sum_{\{\theta,\theta'\}\in E} p_\theta p_{\theta'} \log\frac{1}{p_\theta p_{\theta'}} \le 2(\mathbb{H}_2(p_{\text{E}}) + p_{\text{E}}\log n)$.

Proof of Lemma 6. We can expand the LHS as

$$\begin{aligned}
\text{LHS} &= -\sum_{\theta'} p_{\theta'} \sum_{\theta: r(\theta)\neq r(\theta')} p_\theta(\log p_\theta + \log p_{\theta'}) \\
&= -2\sum_{\theta'} p_{\theta'} \sum_{\theta: r(\theta)\neq r(\theta')} p_\theta \log p_\theta
\end{aligned}$$

$$
\begin{aligned}
&= -2\sum_{y\in\mathcal{Y}}\sum_{\theta':r(\theta')=y} p_{\theta'}\sum_{\theta:r(\theta)\neq y} p_\theta \log p_\theta \\
&= 2\sum_{y\in\mathcal{Y}} p_y(1-p_y)\sum_{\theta:r(\theta)\neq y}\frac{p_\theta}{1-p_y}\left(\log\frac{p_\theta}{1-p_y}+\log(1-p_y)\right) \\
&= -2\sum_{y\in\mathcal{Y}} p_y(1-p_y)\log(1-p_y) + 2\sum_{y\in\mathcal{Y}} p_y(1-p_y)\mathbb{H}\left(\left\{\frac{p_\theta}{(1-p_y)}\right\}_{\theta:r(\theta)\neq y}\right) \qquad \text{(B.3)} \\
&\le 2\sum_{y\in\mathcal{Y}} p_y\mathbb{H}_2(1-p_y) + 2\sum_{y\in\mathcal{Y}} p_y(1-p_y)\mathbb{H}\left(\left\{\frac{p_\theta}{(1-p_y)}\right\}_{\theta:r(\theta)\neq y}\right)
\end{aligned}
$$

Since $\mathbb{H}\left(\left\{\frac{p_\theta}{(1-p_y)}\right\}_{\theta:r(\theta)\neq y}\right) \le \log t \le \log n$, we have

$$
\begin{aligned}
\text{LHS} &\le 2\sum_{y\in\mathcal{Y}} p_y\mathbb{H}_2(1-p_y) + 2\sum_y \underbrace{p_y(1-p_y)\log n}_{p_{\text{E}}\log n} \\
&\overset{\text{Jensen}}{\le} 2\mathbb{H}_2\left(\sum_{y\in\mathcal{Y}} p_y(1-p_y)\right) + 2p_{\text{E}}\log n \\
&= 2\left(\mathbb{H}_2(p_{\text{E}}) + p_{\text{E}}\log n\right).
\end{aligned}
$$

which completes the proof. □

Now, we are ready to state the upper bound $\textsc{Ub}\left(p^{\text{MAP}}_{\text{ERR}}\right)$ and lower bound $\textsc{Lb}\left(p^{\text{MAP}}_{\text{ERR}}\right)$ of f_{AUX}.

Proof of Lemma 2. Clearly, $\sum_{\{\theta,\theta'\}\in E} p_\theta p_{\theta'} \log\frac{1}{p_\theta p_{\theta'}} \ge 0$. By Lemma 5 we get the lower bound:

$$
f_{\text{AUX}}(\psi) \ge 2c\cdot p^{\text{MAP}}_{\text{ERR}}(\psi).
$$

Now assume $p^{\text{MAP}}_{\text{ERR}} \le 1/4$. By Lemma 4 we know $p_{\text{E}} \le 2p^{\text{MAP}}_{\text{ERR}}$, and $\mathbb{H}_2(p_{\text{E}}) \le \mathbb{H}_2\left(2p^{\text{MAP}}_{\text{ERR}}\right) \le 2\mathbb{H}_2\left(p^{\text{MAP}}_{\text{ERR}}\right)$. Combining with Lemma 5 and Lemma 6, we get

$$
\begin{aligned}
f_{\text{AUX}}(\psi) &\le 3c\cdot\left(\mathbb{H}_2\left(p^{\text{MAP}}_{\text{ERR}}\right) + p^{\text{MAP}}_{\text{ERR}}\log n\right) + 2\left(\mathbb{H}_2(p_{\text{E}}) + p_{\text{E}}\log n\right) \\
&\le (3c+4)\cdot\left(\mathbb{H}_2\left(p^{\text{MAP}}_{\text{ERR}}\right) + p^{\text{MAP}}_{\text{ERR}}\log n\right),
\end{aligned}
$$

which completes the proof. □

B.3. Proof of Theorem 1 part 2: Proof of Lemma 3

In this section, we analyze the 1-step gain in the auxiliary function $\Delta_{\text{AUX}}(v \mid \psi)$, of any test $e \in \mathcal{V}$. By the end of this section, we will show that it is lowered bounded by the one-step gain in the EC^2 objective $\Delta_{\text{EC}^2,\psi}(v)$.

Recall that we assume test outcomes are binary for our analysis, and in the following of this section, we assume the outcome x_e of test e is in $\{+,-\}$ instead of $\{0,1\}$, for clarity purposes.

B.3.1. Notations and the intermediate goal

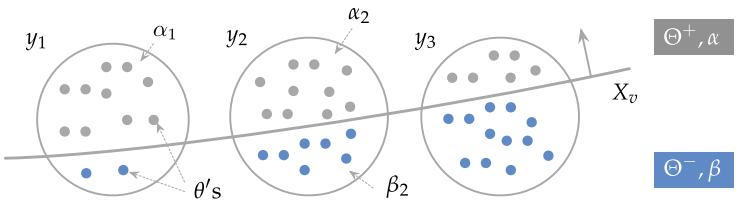


FIG 10. *Performing binary test e on Θ and Y. Dots represent root-causes $\theta \in \text{supp}(\Theta)$, and circles represent values of the target variable $y \in \mathcal{Y}$. The favorable outcome of X_e for the root-causes in gray dots are $+$; the favorable outcome for root-causes in blue dots are $-$. We also illustrate the short-hand notations used in §B.3. They are: p, q (i.e., the posterior probability distribution over Y and Θ), h (i.e., the prior distribution over Y and Θ) and α, β (i.e., the probability mass of gray and brue dots, respectively, before performing test e).*

TABLE 2
Summary of notations introduced for the proof of Lemma 3

h	$\mathbb{P}[\cdot \mid \psi]$, i.e., probability distribution on Θ and Y, before performing test e
h_+, h_-	$\mathbb{P}[X_e = + \mid \psi]$,$\mathbb{P}[X_e = - \mid \psi]$
p_θ, p_y	$\mathbb{P}[\cdot \mid \psi, X_e = +]$, i.e., probability distribution on Θ and Y having observed $X_e = +$
q_θ, q_y	$\mathbb{P}[\cdot \mid \psi, X_e = -]$, i.e., probability distribution on Θ and Y having observed $X_e = -$
Θ^+, Θ^-	set of positive / negative root-causes
Θ_i^+, Θ_i^-	set of positive / negative root-causes associated with target y_i
α, β	total probability mass of positive / negative root-causes
α_i, β_i	probability mass of positive / negative root-causes associated with target y_i
μ_i, ν_i	α_i/α, β_i/β (defined in §B.3.5)
$\theta \nsim \theta'$	$r(\theta) \neq r(\theta')$, i.e., root-causes θ and θ' do not share the same target value

For brevity, we first define a few short-hand notations to simplify our derivation. Let p, q be two distributions on Θ, and $h = h_+ p + h_- q$ be the convex combination of the two, where $h_+, h_- \geq 0$ and $h_+ + h_- = 1$.

In fact, we are using p and q to refer to the posterior distribution over Θ *after* we observe the (noisy) outcome of some binary test e, and use h to refer to the distribution over Θ *before* we perform the test, i.e., $p_\theta \triangleq \mathbb{P}[\theta \mid X_e = +]$, $q_\theta \triangleq \mathbb{P}[\theta \mid X_e = -]$, and $h_\theta \triangleq \mathbb{P}[\theta] = h_+ p_\theta + h_- q_\theta$, where $h_+ = \mathbb{P}[X_e = +]$ and $h_- = \mathbb{P}[X_e = -]$. For $y_i \in \mathcal{Y}$, we use $p_i \triangleq \sum_{\theta: r(\theta) = y_i} p_\theta$ to denote the probability of y_i under distribution p, and use $q_i \triangleq \sum_{\theta: r(\theta) = y_i} q_\theta$ to denote the probability of y_i under distribution q.

Further, given a test e, we define Θ_i^+, Θ_i^- to be the set of root-causes associated with target y_i, whose favorable outcome of test e is $+$ (for Θ_i^+) and $-$ (for Θ_i^-). Formally,

$$\Theta_i^+ \triangleq \{\theta : r(\theta) = y_i \wedge \mathbb{P}[X_e = + \mid \theta] \geq 1/2\}$$
$$\Theta_i^- \triangleq \{\theta : r(\theta) = y_i \wedge \mathbb{P}[X_e = + \mid \theta] < 1/2\}$$

We then define $\Theta^+ \triangleq \bigcup_{i\in\{1,\dots,t\}} \Theta_i^+$, and $\Theta^- \triangleq \bigcup_{i\in\{1,\dots,t\}} \Theta_i^-$, to be the set of "positive" and "negative" root-causes for test e, respectively.

Let α_i, β_i be the probability mass of the root-causes in Θ_i^+ and Θ_i^-, i.e., $\alpha_i \triangleq \sum_{y\in\Theta_i^+} \mathbb{P}[\theta]$, and $\beta_i \triangleq \sum_{y\in\Theta_i^-} \mathbb{P}[\theta]$. We further define $\alpha \triangleq \sum_{y_i\in\mathcal{Y}} \alpha_i = \sum_{\theta\in\Theta^+} \mathbb{P}[\theta]$, and $\beta \triangleq \sum_{y_i\in\mathcal{Y}} \beta_y = \sum_{\theta\in\Theta^-} \mathbb{P}[\theta]$, then clearly we have $\alpha+\beta=1$. See Fig. 10 for illustration.

Now, we assume that test e has error rate ϵ. That is,

$$\forall\theta,\ \ \min\{\mathbb{P}[X_e=+\mid\theta], \mathbb{P}[X_e=-\mid\theta]\} = \epsilon.$$

Then, by definition of h_+, h_-, p_i, q_i, p_θ, q_θ, it is easy to verify that

$$\begin{aligned}
&h_+ = \alpha\bar{\epsilon}+\beta\epsilon, \quad h_- = \alpha\epsilon+\beta\bar{\epsilon}\\
&p_i = \frac{\alpha_i\bar{\epsilon}+\beta_i\epsilon}{h_+}, \quad q_i = \frac{\alpha_i\epsilon+\beta_i\bar{\epsilon}}{h_-}\\
&p_\theta = \frac{h_\theta\bar{\epsilon}}{h_+}, \quad q_\theta = \frac{h_\theta\epsilon}{h_-}, \qquad \text{if } \theta\in\Theta_i^+\\
&p_\theta = \frac{h_\theta\epsilon}{h_+}, \quad q_\theta = \frac{h_\theta\bar{\epsilon}}{h_-}, \qquad \text{if } \theta\in\Theta_i^-
\end{aligned} \tag{B.4}$$

For the convenience of readers, we summarize the notations provided above in Table 2.

Given root-causes θ and θ', we use $\theta \nsim \theta'$ to denote that the values of the target variable Y associated with root-causes θ and θ' are different, i.e., $r(\theta)\neq r(\theta')$.

We can rewrite the auxiliary function (as defined in Equation (4.1)) as follows:

$$f_{\text{AUX}} = \sum_{\theta\nsim\theta'} h_\theta h_{\theta'} \log\frac{1}{h_\theta h_{\theta'}} + c\sum_{y_i\in\mathcal{Y}} \mathbb{H}_2(h_i).$$

If by performing test e we observe $X_e=+$, we have

$$f_{\text{AUX}}((e,+)) = \sum_{\theta\nsim\theta'} p_\theta p_{\theta'} \log\frac{1}{p_\theta p_{\theta'}} + c\sum_{y_i\in\mathcal{Y}} \mathbb{H}_2(p_i)$$

otherwise, if we observe $X_e=-$,

$$f_{\text{AUX}}((e,-)) = \sum_{\theta\nsim\theta'} q_\theta q_{\theta'} \log\frac{1}{q_\theta q_{\theta'}} + c\sum_{y_i\in\mathcal{Y}} \mathbb{H}_2(q_i)$$

Therefore, the expected gain (i.e., $\Delta_{\text{AUX}}(v\mid\psi)$) of performing test e is,

$$\Delta_{\text{AUX}} = \overbrace{\sum_{\theta\nsim\theta'} h_\theta h_{\theta'} \log\frac{1}{h_\theta h_{\theta'}} - \left(h_+\sum_{\theta\nsim\theta'} p_\theta p_{\theta'}\log\frac{1}{p_\theta p_{\theta'}} + h_-\sum_{\theta\nsim\theta'} q_\theta q_{\theta'}\log\frac{1}{q_\theta q_{\theta'}}\right)}^{①}$$
$$+ \underbrace{c\left(\sum_{y_i\in\mathcal{Y}}\mathbb{H}_2(h_i) - \left(h_+\sum_{y_i\in\mathcal{Y}}\mathbb{H}_2(p_i) + h_-\sum_{y_i\in\mathcal{Y}}\mathbb{H}_2(q_i)\right)\right)}_{②} \tag{B.5}$$

In the following, we derive lower bounds for the above two terms respectively.

B.3.2. A lower bound on term 1

Let $g_{\theta,\theta'} \triangleq h_+ p_\theta p_{\theta'} + h_- q_\theta q_{\theta'}$. Then, we can rewrite Term ❶ as,

$$\text{Term } ❶ = \underbrace{\sum_{\theta \nsim \theta'} h_\theta h_{\theta'} \log \frac{1}{h_\theta h_{\theta'}} - \sum_{\theta \nsim \theta'} g_{\theta,\theta'} \log \frac{1}{g_{\theta,\theta'}}}_{\text{Part 1}} + \underbrace{\sum_{\theta \nsim \theta'} g_{\theta,\theta'} \log \frac{1}{g_{\theta,\theta'}} - \left(h_+ \sum_{\theta \nsim \theta'} p_\theta p_{\theta'} \log \frac{1}{p_\theta p_{\theta'}} + h_- \sum_{\theta \nsim \theta'} q_\theta q_{\theta'} \log \frac{1}{q_\theta q_{\theta'}} \right)}_{\text{Part 2}} \tag{B.6}$$

Part 1. We first provide a lower bound for part 1 of Equation (B.6).

Notice that for the concave function $f(x) = x \log \frac{1}{x}$ and $\delta < x$, it holds that $f(x) - f(x-\delta) \geq \delta \frac{\partial f}{\partial x}\big|_x = \delta(\log \frac{1}{x} - 1)$, then we get

$$\sum_{\theta \nsim \theta'} h_\theta h_{\theta'} \log \frac{1}{h_\theta h_{\theta'}} - \sum_{\theta \nsim \theta'} g_{\theta,\theta'} \log \frac{1}{g_{\theta,\theta'}} \geq \sum_{\theta \nsim \theta'} (h_\theta h_{\theta'} - g_{\theta,\theta'}) \left(\log \frac{1}{h_\theta h_{\theta'}} - 1 \right)$$

Further, observe

$$\begin{aligned} h_\theta h_{\theta'} - g_{\theta,\theta'} &= (h_+ p_\theta + h_- q_\theta)(h_+ p_{\theta'} + h_- q_{\theta'}) - (h_+ p_\theta p_{\theta'} + h_- q_\theta q_{\theta'}) \\ &= (h_+ p_\theta + h_- q_\theta)(p_{\theta'} + q_{\theta'} - h_- p_{\theta'} - h_+ q_{\theta'}) - (h_+ p_\theta p_{\theta'} + h_- q_\theta q_{\theta'}) \\ &= h_+ h_- p_{\theta'} q_\theta - h_+ h_- p_{\theta'} p_\theta + h_+ h_- p_\theta q_{\theta'} - h_- h_+ q_{\theta'} q_\theta \\ &= -h_+ h_- (p_\theta - q_\theta)(p_{\theta'} - q_{\theta'}) \end{aligned}$$

Combining the above two equations gives us

$$\text{Part 1} \geq \sum_{\theta \nsim \theta'} -h_+ h_- (p_\theta - q_\theta)(p_{\theta'} - q_{\theta'}) \left(\log \frac{1}{h_\theta h_{\theta'}} - 1 \right)$$

For any root-cause pair $\{\theta, \theta'\}$ with $\theta \nsim \theta'$, and binary test e, there are only 4 possible combinations in terms of the root-causes' favorable outcomes. Namely,

1. Both θ and θ' maps x to $+$, i.e., $\theta \in \Theta^+ \wedge \theta' \in \Theta^+$.
   We define such set of root-cause pairs with positive favorable outcomes as $U_{(+,+)} \triangleq \{\{\theta, \theta'\} : \theta \in \Theta^+ \wedge \theta' \in \Theta^+\}$ (For other cases, we define $U_{(-,-)}$, $U_{(+,-)}$, $U_{(-,+)}$ in a similar way).
   In this case, we have

$$\begin{aligned} &\sum_{\{\theta,\theta'\} \in U_{(+,+)}} -h_+ h_- (p_\theta - q_\theta)(p_{\theta'} - q_{\theta'}) \left(\log \frac{1}{h_\theta h_{\theta'}} - 1 \right) \\ \overset{\text{Eq (B.4)}}{=} &\sum_{\{\theta,\theta'\} \in U_{(+,+)}} -h_+ h_- \left(\frac{h_\theta \bar{\epsilon}}{h_+} - \frac{h_\theta \epsilon}{h_-} \right) \left(\frac{h_{\theta'} \bar{\epsilon}}{h_+} - \frac{h_{\theta'} \epsilon}{h_-} \right) \left(\log \frac{1}{h_\theta h_{\theta'}} - 1 \right) \end{aligned}$$

$$
\begin{aligned}
=&h_+h_-\left(\frac{h_-\bar{\epsilon}-h_+\bar{\epsilon}}{h_+h_-}\right)^2\sum_{\{\theta,\theta'\}\in U_{(+,+)}}-h_\theta h_{\theta'}\left(\log\frac{1}{h_\theta h_{\theta'}}-1\right)\\
=&\frac{\beta^2\left(1-2\epsilon\right)^2}{h_+h_-}\sum_{\{\theta,\theta'\}\in U_{(+,+)}}-h_\theta h_{\theta'}\left(\log\frac{1}{h_\theta h_{\theta'}}-1\right)\\
=&\frac{\beta^2\left(1-2\epsilon\right)^2}{h_+h_-}\sum_{\{\theta,\theta'\}\in U_{(+,+)}}\left(-2h_\theta h_{\theta'}\log\frac{1}{h_\theta}+h_\theta h_{\theta'}\right)\\
=&\frac{\beta^2\left(1-2\epsilon\right)^2}{h_+h_-}\left(\sum_{y_i\in\mathcal{Y}}(\alpha-\alpha_i)\sum_{\theta\in\Theta_i^+}-2h_\theta\log\frac{1}{h_\theta}+\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)\right)\\
=&\frac{\left(1-2\epsilon\right)^2}{h_+h_-}\left(-2\beta^2\sum_{y_i\in\mathcal{Y}}(\alpha-\alpha_i)\sum_{\theta\in\Theta_i^+}h_\theta\log\frac{1}{h_\theta}+\beta^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)\right)
\end{aligned}
$$

2. Both θ and θ' maps x to $-$. Similarly, we get

$$
\begin{aligned}
&\sum_{\{\theta,\theta'\}\in U_{(-,-)}}-h_+h_-(p_\theta-q_\theta)(p_{\theta'}-q_{\theta'})\left(\log\frac{1}{h_\theta h_{\theta'}}-1\right)\\
=&\frac{\left(1-2\epsilon\right)^2}{h_+h_-}\left(-2\alpha^2\sum_{y_i\in\mathcal{Y}}(\beta-\beta_i)\sum_{\theta\in\Theta_i^-}h_\theta\log\frac{1}{h_\theta}+\alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)\right)
\end{aligned}
$$

3. θ maps x to $+$, θ' maps x to $-$. We have

$$
\begin{aligned}
&\sum_{(\theta,\theta')\in U_{(+,-)}}\frac{h_+h_-}{2}\frac{\left(p_\theta p_{\theta'}-q_\theta q_{\theta'}\right)^2}{p_\theta p_{\theta'}+q_\theta q_{\theta'}}\\
\geq&\sum_{(\theta,\theta')\in U_{(+,+)}}\frac{h_+h_-}{2}\left(\sqrt{\frac{h_\theta\bar{\epsilon}}{h_+}\frac{h_{\theta'}\epsilon}{h_+}}-\sqrt{\frac{h_\theta\epsilon}{h_-}\frac{h_{\theta'}\bar{\epsilon}}{h_-}}\right)^2\\
=&\sum_{(\theta,\theta')\in U_{(+,+)}}\frac{h_+h_-}{2}h_\theta h_{\theta'}\epsilon\bar{\epsilon}\left(\frac{1}{h_+}-\frac{1}{h_-}\right)^2\\
=&\frac{\left(1-2\epsilon\right)^2}{2h_+h_-}\epsilon\bar{\epsilon}(\alpha-\beta)^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)
\end{aligned}
$$

4. θ maps x to $-$, θ' maps x to $+$. By symmetry we have

$$
\begin{aligned}
&\sum_{(\theta,\theta')\in U_{(-,+)}}-h_+h_-(p_\theta-q_\theta)(p_{\theta'}-q_{\theta'})\left(\log\frac{1}{h_\theta h_{\theta'}}-1\right)\\
=&\sum_{(\theta,\theta')\in U_{(+,-)}}-h_+h_-(p_\theta-q_\theta)(p_{\theta'}-q_{\theta'})\left(\log\frac{1}{h_\theta h_{\theta'}}-1\right)
\end{aligned}
$$

Combining the above four equations, we obtain a lower bound on Part 1:

$$\begin{aligned}
\text{Part 1} \geq \frac{(1-2\epsilon)^2}{h_+h_-}\Bigg(&-2\beta^2\sum_{y_i\in\mathcal{Y}}(\alpha-\alpha_i)\sum_{\theta\in\Theta_i^+}h_\theta\log\frac{1}{h_\theta}+\beta^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)\\
&-2\alpha^2\sum_{y_i\in\mathcal{Y}}(\beta-\beta_i)\sum_{\theta\in\Theta_i^-}h_\theta\log\frac{1}{h_\theta}+\alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)\\
&+2\alpha\beta\sum_{y_i\in\mathcal{Y}}(\beta-\beta_i)\sum_{\theta\in\Theta_i^+}h_\theta\log\frac{1}{h_\theta}\\
&+2\alpha\beta\sum_{y_i\in\mathcal{Y}}(\alpha-\alpha_i)\sum_{\theta\in\Theta_i^-}h_\theta\log\frac{1}{h_\theta}-2\alpha\beta\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\Bigg)\\
=\frac{(1-2\epsilon)^2}{h_+h_-}\Bigg(&\Bigg(2\alpha\beta\sum_{y_i\in\mathcal{Y}}(\beta-\beta_i)-2\beta^2\sum_{y_i\in\mathcal{Y}}(\alpha-\alpha_i)\Bigg)\sum_{\theta\in\Theta_i^+}h_\theta\log\frac{1}{h_\theta}\\
&+\Bigg(2\alpha\beta\sum_{y_i\in\mathcal{Y}}(\alpha-\alpha_i)-2\alpha^2\sum_{y_i\in\mathcal{Y}}(\beta-\beta_i)\Bigg)\sum_{\theta\in\Theta_i^-}h_\theta\log\frac{1}{h_\theta}\\
&+\beta^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)+\alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)-2\alpha\beta\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\Bigg)\\
=\frac{(1-2\epsilon)^2}{h_+h_-}\cdot\Bigg(&2\sum_{y_i\in\mathcal{Y}}\beta(\beta\alpha_i-\alpha\beta_i)\sum_{\theta\in\Theta_i^+}h_\theta\log\frac{1}{h_\theta}\\
&+2\sum_{y_i\in\mathcal{Y}}\alpha(\alpha\beta_i-\beta\alpha_i)\sum_{\theta\in\Theta_i^-}h_\theta\log\frac{1}{h_\theta}-\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)^2\Bigg)\\
=\frac{(1-2\epsilon)^2}{h_+h_-}\cdot\Bigg(&2\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)\Bigg(\beta\alpha_i\sum_{\theta\in\Theta_i^+}\frac{h_\theta}{\alpha_i}\log\frac{1}{h_\theta}\\
&-\alpha\beta_i\sum_{\theta\in\Theta_i^-}\frac{h_\theta}{\beta_i}\log\frac{1}{h_\theta}\Bigg)-\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)^2\Bigg)
\end{aligned}\tag{B.7}$$

Part 2. Next, we will provide a lower bound on Part 2 of Equation (B.6). By definition, we have

$$\begin{aligned}
\text{Part 2}=&\sum_{\theta\nsim\theta'}(h_+p_\theta p_{\theta'}+h_-q_\theta q_{\theta'})\log\frac{1}{h_+p_\theta p_{\theta'}+h_-q_\theta q_{\theta'}}\\
&-\Bigg(h_+\sum_{\theta\nsim\theta'}p_\theta p_{\theta'}\log\frac{1}{p_\theta p_{\theta'}}+h_-\sum_{\theta\nsim\theta'}q_\theta q_{\theta'}\log\frac{1}{q_\theta q_{\theta'}}\Bigg)\\
\overset{(a)}{\geq}&\frac{h_+h_-}{2}\sum_{\theta\nsim\theta'}\frac{(p_\theta p_{\theta'}-q_\theta q_{\theta'})^2}{p_\theta p_{\theta'}+q_\theta q_{\theta'}}
\end{aligned}$$

Hereby, step (a) is due to the strong concavity[6] of $f(x) = x \log \frac{1}{x}$.

Similarly with the analysis of Part 1, we consider the four sets of $\{\theta, \theta'\}$ pairs:

1. $\{\theta, \theta'\} \in U_{(+,+)}$: both θ and θ' maps x to $+$.
   In this case, we have

$$\begin{aligned}
&\sum_{(\theta,\theta')\in U_{(+,+)}} \frac{h_+h_-}{2}\frac{\left(p_\theta p_{\theta'} - q_\theta q_{\theta'}\right)^2}{p_\theta p_{\theta'} + q_\theta q_{\theta'}} \\
\geq &\sum_{(\theta,\theta')\in U_{(+,+)}} \frac{h_+h_-}{2}\left(\sqrt{p_\theta p_{\theta'}} - \sqrt{q_\theta q_{\theta'}}\right)^2 \\
\overset{\text{Eq (B.4)}}{=} &\sum_{(\theta,\theta')\in U_{(+,+)}} \frac{h_+h_-}{2}\left(\sqrt{\frac{h_\theta\bar{\epsilon}}{h_+}\frac{h_{\theta'}\bar{\epsilon}}{h_+}} - \sqrt{\frac{h_\theta\epsilon}{h_-}\frac{h_{\theta'}\epsilon}{h_-}}\right)^2 \\
= &\sum_{(\theta,\theta')\in U_{(+,+)}} \frac{h_+h_-}{2} h_\theta h_{\theta'}\left(\frac{\bar{\epsilon}}{h_+} - \frac{\epsilon}{h_-}\right)^2 \\
= &\sum_{(\theta,\theta')\in U_{(+,+)}} \frac{h_+h_-}{2} h_\theta h_{\theta'}\frac{\beta^2\left(1-2\epsilon\right)^2}{(h_+h_-)^2} \\
= &\frac{\left(1-2\epsilon\right)^2}{2h_+h_-}\beta^2 \sum_{y_i\in\mathcal{Y}} \alpha_i(\alpha - \alpha_i)
\end{aligned}$$

2. $(\theta, \theta') \in U_{(-,-)}$. Similarly, we get

$$\sum_{(\theta,\theta')\in U_{(-,-)}} \frac{h_+h_-}{2}\frac{\left(p_\theta p_{\theta'} - q_\theta q_{\theta'}\right)^2}{p_\theta p_{\theta'} + q_\theta q_{\theta'}} \geq \frac{\left(1-2\epsilon\right)^2}{2h_+h_-}\alpha^2 \sum_{y_i\in\mathcal{Y}} \beta_i(\beta - \beta_i)$$

3. $(\theta, \theta') \in U_{(+,-)}$: θ maps x to $+$, θ' maps x to $-$. We have

$$\begin{aligned}
&\sum_{(\theta,\theta')\in U_{(+,-)}} \frac{h_+h_-}{2}\frac{\left(p_\theta p_{\theta'} - q_\theta q_{\theta'}\right)^2}{p_\theta p_{\theta'} + q_\theta q_{\theta'}} \\
\geq &\sum_{(\theta,\theta')\in U_{(+,+)}} \frac{h_+h_-}{2}\left(\sqrt{\frac{h_\theta\bar{\epsilon}}{h_+}\frac{h_{\theta'}\epsilon}{h_+}} - \sqrt{\frac{h_\theta\epsilon}{h_-}\frac{h_{\theta'}\bar{\epsilon}}{h_-}}\right)^2 \\
= &\sum_{(\theta,\theta')\in U_{(+,+)}} \frac{h_+h_-}{2} h_\theta h_{\theta'}\epsilon\bar{\epsilon}\left(\frac{1}{h_+} - \frac{1}{h_-}\right)^2 \\
= &\frac{\left(1-2\epsilon\right)^2}{2h_+h_-}\epsilon\bar{\epsilon}(\alpha - \beta)^2 \sum_{y_i\in\mathcal{Y}} \alpha_i(\beta - \beta_i)
\end{aligned}$$

[6] If f is strongly concave, then for $t \in [0,1]$, it holds that $f(tx+(1-t)y) - tf(x) - (1-t)f(y) \geq \frac{t(1-t)}{2}m(x-y)^2$, where $m = \min\left(|f''(x)|, |f''(y)|\right)$.

4. $(\theta,\theta')\in U_{(-,+)}$: θ maps x to $-$, θ' maps x to $+$. By symmetry we have

$$\sum_{(\theta,\theta')\in U_{(+,-)}}\frac{h_+h_-}{2}\frac{\left(p_\theta p_{\theta'}-q_\theta q_{\theta'}\right)^2}{p_\theta p_{\theta'}+q_\theta q_{\theta'}}\geq\frac{(1-2\epsilon)^2}{2h_+h_-}\epsilon\bar{\epsilon}(\alpha-\beta)^2\sum_{y_i\in\mathcal{Y}}\beta_i(\alpha-\alpha_i)$$

Combining the above four equations, we obtain a lower bound on Part 2:

$$\begin{aligned}&\text{Part 2}\\&\geq\sum_{(\theta,\theta')\in U_{(+,+)}}\frac{h_+h_-}{2}\frac{\left(p_\theta p_{\theta'}-q_\theta q_{\theta'}\right)^2}{p_\theta p_{\theta'}+q_\theta q_{\theta'}}+\sum_{(\theta,\theta')\in U_{(-,-)}}\frac{h_+h_-}{2}\frac{\left(p_\theta p_{\theta'}-q_\theta q_{\theta'}\right)^2}{p_\theta p_{\theta'}+q_\theta q_{\theta'}}\\&+\sum_{(\theta,\theta')\in U_{(+,-)}}\frac{h_+h_-}{2}\frac{\left(p_\theta p_{\theta'}-q_\theta q_{\theta'}\right)^2}{p_\theta p_{\theta'}+q_\theta q_{\theta'}}+\sum_{(\theta,\theta')\in U_{(-,+)}}\frac{h_+h_-}{2}\frac{\left(p_\theta p_{\theta'}-q_\theta q_{\theta'}\right)^2}{p_\theta p_{\theta'}+q_\theta q_{\theta'}}\\&=\tfrac{(1-2\epsilon)^2}{2h_+h_-}\cdot\left(\beta^2\textstyle\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)+\alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)+2\epsilon\bar{\epsilon}(\alpha-\beta)^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\right)\end{aligned}\tag{B.8}$$

B.3.3. A lower bound on term 2

Now we move on to analyze Term ❷ of Equation (B.6). By strong concavity of $f(x)=x\log\frac{1}{x}+(1-x)\log\frac{1}{1-x}$, we obtain

$$\begin{aligned}\text{Term ❷}&=c\sum_{y_i\in\mathcal{Y}}\Bigg(h_i\log\frac{1}{h_i}+(1-h_i)\log\frac{1}{1-h_i}\\&\qquad-h_+\left(p_i\log\frac{1}{p_i}+(1-p_i)\log\frac{1}{1-p_i}\right)\\&\qquad-h_-\left(q_i\log\frac{1}{q_i}+(1-q_i)\log\frac{1}{1-q_i}\right)\Bigg)\\&\overset{\text{footnote 6}}{\geq}\frac{c\cdot h_+h_-}{2}\sum_{y_i\in\mathcal{Y}}\frac{\left(p_i-q_i\right)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}\end{aligned}$$

Plugging in the definition of p_i, q_i from Equation (B.4), we get

$$\begin{aligned}\text{Term ❷}&=\frac{c\cdot h_+h_-}{2}\sum_{y_i\in\mathcal{Y}}\left(\frac{\alpha_i\bar{\epsilon}+\beta_i\epsilon}{h_+}-\frac{\alpha_i\epsilon+\beta_i\bar{\epsilon}}{h_-}\right)^2\frac{1}{\max\{p_i(1-p_i),q_i(1-q_i)\}}\\&=\frac{c}{2h_+h_-}\sum_{y_i\in\mathcal{Y}}\frac{\left((\alpha\epsilon+\beta\bar{\epsilon})(\alpha_i\bar{\epsilon}+\beta_i\epsilon)-(\alpha\bar{\epsilon}+\beta\epsilon)(\alpha_i\epsilon+\beta_i\bar{\epsilon})\right)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}\\&=\frac{c}{2h_+h_-}\sum_{y_i\in\mathcal{Y}}\frac{\left(\alpha\beta_i\epsilon^2+\beta\alpha_i\bar{\epsilon}^2-\alpha\beta_i\bar{\epsilon}^2-\beta\alpha_i\epsilon^2\right)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}\\&=\frac{c(1-2\epsilon)^2}{2h_+h_-}\sum_{y_i\in\mathcal{Y}}\frac{\left(\beta\alpha_i-\alpha\beta_i\right)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}\end{aligned}\tag{B.9}$$

B.3.4. A combined lower bound for Δ_{AUX}

Now, combining Equation (B.7), (B.8), and (B.9), we can get a lower bound for Δ_{AUX}:

$$\begin{aligned}\Delta_{\text{AUX}} \geq \frac{(1-2\epsilon)^2}{h_+h_-}\Bigg(&2\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)\\&\times\left(\beta\alpha_i\sum_{\theta\in\Theta_i^+}\frac{h_\theta}{\alpha_i}\log\frac{1}{h_\theta}-\alpha\beta_i\sum_{\theta\in\Theta_i^-}\frac{h_\theta}{\beta_i}\log\frac{1}{h_\theta}\right)-\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)^2\Bigg)\\&+\frac{(1-2\epsilon)^2}{2h_+h_-}\Bigg(\beta^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)\\&+\alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)+2\epsilon\bar{\epsilon}(\alpha-\beta)^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\Bigg)\\&+\frac{c(1-2\epsilon)^2}{2h_+h_-}\sum_{y_i\in\mathcal{Y}}\frac{(\beta\alpha_i-\alpha\beta_i)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}\end{aligned}\tag{B.10}$$

We can rewrite Equation (B.10) as

$$\begin{aligned}\Delta_{\text{AUX}} \geq \frac{(1-2\epsilon)^2}{4h_+h_-}\underbrace{\Bigg(\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)^2+\beta^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)+\alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)+2\epsilon\bar{\epsilon}(\alpha-\beta)^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\Bigg)}_{\text{LB1}}\\+\frac{(1-2\epsilon)^2}{4h_+h_-}\underbrace{\Bigg(\beta^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)+\alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)+2\epsilon\bar{\epsilon}(\alpha-\beta)^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)}_{}\\\underbrace{+2c\sum_{y_i\in\mathcal{Y}}\frac{(\beta\alpha_i-\alpha\beta_i)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}-5\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)^2+8\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)\left(\beta\alpha_i\sum_{\theta\in\Theta_i^+}\frac{h_\theta}{\alpha_i}\log\frac{1}{h_\theta}-\alpha\beta_i\sum_{\theta\in\Theta_i^-}\frac{h_\theta}{\beta_i}\log\frac{1}{h_\theta}\right)\Bigg)}_{\text{LB2}}\end{aligned}\tag{B.11}$$

B.3.5. Connecting Δ_{AUX} *with* Δ_{EC^2}

Next, we will show that term LB1 is lower-bounded by a factor of Δ_{EC^2} (i.e., $\Delta_{\text{EC}^2,\psi}(v)$), while LB2 cannot be too much less than 0. Concretely, we will show

- LB1 $\geq \frac{1}{16}(1-2\epsilon)^2 \Delta_{\mathrm{EC}^2}$, and
- LB2 $\geq -2t(1-2\epsilon)^2 \eta$, for $\eta \in (0,1)$.

At the end of this subsection, we will combine the above results to connect $\Delta_{\mathrm{AUX}}(v \mid \psi)$ with $\Delta_{\mathrm{EC}^2,\psi}(v)$ (See Equation (B.18)).

LB1 VS. Δ_{EC^2}. We expand the EC^2 gain $\Delta_{\mathrm{EC}^2,\psi}(v)$ as

$$\begin{aligned}\Delta_{\mathrm{EC}^2} &= \sum_{y_i\in\mathcal{Y}}(\alpha_i+\beta_i)(1-\alpha_i-\beta_i) - \alpha\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i) - \beta\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)\\ &= \beta\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i) + \alpha\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i) + 2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i) \qquad \text{(B.12)}\end{aligned}$$

Define

$$\begin{cases} \textcircled{*} \triangleq \frac{16h_+h_-}{(1-2\epsilon)^2}\cdot \mathrm{LB1} \\ \qquad = 4\Big(\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)^2 + \beta^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i) + \alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i) \\ \qquad\quad +2\epsilon\bar{\epsilon}(\alpha-\beta)^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\Big) \\ \textcircled{\dagger} \triangleq h_+h_-\Delta_{\mathrm{EC}^2} \\ \qquad = \big(\epsilon\bar{\epsilon}(\alpha-\beta)^2+\alpha\beta\big)\Big(\beta\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i) + \alpha\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i) \\ \qquad\quad +2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\Big)\end{cases}$$

To bound LB1 against $\frac{1}{16}(1-2\epsilon)^2\Delta_{\mathrm{EC}^2}$, it suffices to show $\textcircled{*} \geq \textcircled{\dagger}$. To prove the above inequality, we consider the following two cases:

1. $\epsilon\bar{\epsilon}(\alpha-\beta)^2 \leq \alpha\beta$. In this case, we have $\epsilon\bar{\epsilon}(\alpha-\beta)^2+\alpha\beta \leq 2\alpha\beta$. Then,

$$\begin{aligned}&\frac{\textcircled{*}-\textcircled{\dagger}}{2}\\ &\geq \frac{\textcircled{*}}{2} - \alpha\beta\left(\beta\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i) + \alpha\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i) + 2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\right)\\ &\geq \beta^2(1+\beta)\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i) + \alpha^2(1+\alpha)\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)\\ &\quad + \sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)^2 - 2\alpha\beta\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\\ &\geq \sum_{y_i\in\mathcal{Y}}\big(\beta^2\alpha_i(\alpha-\alpha_i) + \alpha^2\beta_i(\beta-\beta_i) + (\beta\alpha_i-\alpha\beta_i)^2 - 2\alpha\beta\alpha_i(\beta-\beta_i)\big)\\ &= 0\end{aligned}$$

2. $\epsilon\bar{\epsilon}(\alpha-\beta)^2 > \alpha\beta$. W.l.o.g., we assume $\beta \leq \alpha \leq 1$. By $\alpha+\beta=1$ we get $2\alpha \geq 1$.

Observe the fact that

$$\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)^2=-\beta^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)\\ -\alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)+2\alpha\beta\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\geq 0$$

Rearranging the terms in the above inequality, we get

$$\beta\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)\leq 2\alpha\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\\ \leq 2(\alpha\beta-\sum_{y_i\in\mathcal{Y}}\alpha_i\beta_i)=2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i) \tag{B.13}$$

Hence,

$$\text{†}\leq 2\epsilon\bar{\epsilon}(\alpha-\beta)^2\left(\beta\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)+\alpha\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)+2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\right)$$

$$\overset{\text{(B.13)}}{\leq} 2\epsilon\bar{\epsilon}(\alpha-\beta)^2\left(\alpha\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)+4\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\right)$$

$$\overset{2\alpha\geq 1}{\leq} 2\epsilon\bar{\epsilon}(\alpha-\beta)^2\left(2\alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)+4\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)\right)$$

$$\overset{\epsilon\bar{\epsilon}(\alpha-\beta)^2\leq 1}{\leq} 4\left(2\epsilon\bar{\epsilon}(\alpha-\beta)^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\beta-\beta_i)+\alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)\right)$$

$$\leq \text{*}$$

Therefore, we get

$$\text{LB1}\geq\frac{1}{16}(1-2\epsilon)^2\,\Delta_{\text{EC}^2} \tag{B.14}$$

A lower bound on LB2. In the following, we will analyze LB2.

$$\text{LB2}\geq\frac{(1-2\epsilon)^2}{4h_+h_-}\left(\beta^2\sum_{y_i\in\mathcal{Y}}\alpha_i(\alpha-\alpha_i)+\alpha^2\sum_{y_i\in\mathcal{Y}}\beta_i(\beta-\beta_i)-5\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)^2\right.\\ +2c_2\sum_{y_i\in\mathcal{Y}}\frac{(\beta\alpha_i-\alpha\beta_i)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}\\ +8\sum_{y_i\in\mathcal{Y}}(\beta\alpha_i-\alpha\beta_i)\left(\beta\alpha_i\sum_{\theta\in\Theta_i^+}\frac{h_\theta}{\alpha_i}\log\frac{\alpha_i}{h_\theta}+\beta\alpha_i\log\frac{1}{\alpha_i}\right.\\ \left.\left.-\alpha\beta_i\sum_{\theta\in\Theta_i^-}\frac{h_\theta}{\beta_i}\log\frac{\beta_i}{h_\theta}-\alpha\beta_i\log\frac{1}{\beta_i}\right)\right)$$

For brevity, define $\mu_i \triangleq \alpha_i/\alpha$, and $\nu_i \triangleq \beta_i/\beta$. We can simplify the above equation as

$$\begin{aligned}\mathrm{LB2} \geq \frac{\alpha^2\beta^2\left(1-2\epsilon\right)^2}{4h_+h_-}\sum_{y_i\in\mathcal{Y}}\Bigg(&\mu_i(1-\mu_i)+\nu_i(1-\nu_i)\\&-5(\mu_i-\nu_i)^2+\frac{2c_2\left(\mu_i-\nu_i\right)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}\\&+8(\mu_i-\nu_i)\Bigg(\mu_i\sum_{\theta\in\Theta_i^+}\frac{h_\theta}{\alpha_i}\log\frac{\alpha_i}{h_\theta}+\mu_i\log\frac{1}{\mu_i\alpha}\\&-\nu_i\sum_{\theta\in\Theta_i^-}\frac{h_\theta}{\beta_i}\log\frac{\beta_i}{h_\theta}-\nu_i\log\frac{1}{\nu_i\beta}\Bigg)\Bigg)\end{aligned}\tag{B.15}$$

Denote the summand on the RHS of the above equation as $\mathrm{LB2}_i$. If for any $y_i \in \mathcal{Y}$ we can lower bound $\mathrm{LB2}_i$, we can then bound the whole sum. Fix i. W.l.o.g., we assume $\mu_i \geq \nu_i$. Then

$$\begin{aligned}\mathrm{LB2}_i &\triangleq \mu_i(1-\mu_i)+\nu_i(1-\nu_i)-5(\mu_i-\nu_i)^2+\frac{2c\left(\mu_i-\nu_i\right)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}\\&+8(\mu_i-\nu_i)\Bigg(\underbrace{\mu_i\sum_{\theta\in\Theta_i^+}\frac{h_\theta}{\alpha_i}\log\frac{\alpha_i}{h_\theta}+\mu_i\log\frac{1}{\mu_i\alpha}}_{\geq 0}-\nu_i\sum_{\theta\in\Theta_i^-}\frac{h_\theta}{\beta_i}\log\frac{\beta_i}{h_\theta}-\nu_i\log\frac{1}{\nu_i\beta}\Bigg)\\&\geq \mu_i(1-\mu_i)+\nu_i(1-\nu_i)-5(\mu_i-\nu_i)^2+\frac{2c\left(\mu_i-\nu_i\right)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}\\&\quad-8(\mu_i-\nu_i)\Bigg(\nu_i\underbrace{\sum_{\theta\in\Theta_i^-}\frac{h_\theta}{\beta_i}\log\frac{\beta_i}{h_\theta}}_{\leq\log n}+\nu_i\log\frac{1}{\nu_i}+\nu_i\log\frac{1}{\beta}\Bigg)\\&\geq \mu_i(1-\mu_i)+\nu_i(1-\nu_i)-5(\mu_i-\nu_i)^2-8(\mu_i-\nu_i)\left(\nu_i\log\frac{n}{\beta}+\nu_i\log\frac{1}{\nu_i}\right)\\&\quad+\frac{2c\left(\mu_i-\nu_i\right)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}\end{aligned}$$

In order to put a lower bound on the above terms, we first need to lower bound the term involving $\frac{(\mu_i-\nu_i)^2}{\max\{p_i(1-p_i),q_i(1-q_i)\}}$. Notice that $p_i = \frac{\alpha_i+\beta_i\epsilon/\bar{\epsilon}}{\alpha+\beta\epsilon/\bar{\epsilon}}$, and $p_i = \frac{\alpha_i\epsilon/\bar{\epsilon}+\beta_i}{\alpha\epsilon/\bar{\epsilon}+\beta}$. Therefore, $\min\{\mu_i,\nu_i\} \leq p_i, q_i \leq \max\{\mu_i,\nu_i\}$.

We check three different cases:

- $\mu_i \geq \nu_i \geq 1/2$, or $\nu_i \leq \mu_i \leq 1/2$.
  In this case, $\max\{p_i(1-p_i),q_i(1-q_i)\} \leq \max\{\mu_i(1-\mu_i),\nu_i(1-\nu_i)\}$. Therefore,

$$\begin{aligned}
\mathrm{LB2}_i &\geq -5(\mu_i-\nu_i)^2 - 8(\mu_i-\nu_i)\left(\nu_i\log\frac{n}{\beta}+\nu_i\log\frac{1}{\nu_i}\right)\\
&\quad + \frac{2c\,(\mu_i-\nu_i)^2}{\max\{\mu_i(1-\mu_i),\nu_i(1-\nu_i)\}} + \mu_i(1-\mu_i)+\nu_i(1-\nu_i)\\
&\geq -5(\mu_i-\nu_i)^2 - 8(\mu_i-\nu_i)\left(\nu_i\log\frac{n}{\beta}+\nu_i\log\frac{1}{\nu_i}\right)\\
&\quad + \frac{2c\,(\mu_i-\nu_i)^2}{\max\{\mu_i(1-\mu_i),\nu_i(1-\nu_i)\}} + \max\{\mu_i(1-\mu_i),\nu_i(1-\nu_i)\}\\
&\geq -5(\mu_i-\nu_i)^2 - 8(\mu_i-\nu_i)\left(\nu_i\log\frac{n}{\beta}+\nu_i\log\frac{1}{\nu_i}\right) + 2\sqrt{2c}(\mu_i-\nu_i)\\
&\overset{\mu_i-\nu_i\leq 1/2}{\geq} (\mu_i-\nu_i)\left(2\sqrt{2c}-5/2-8\left(\nu_i\log\frac{n}{\beta}+\nu_i\log\frac{1}{\nu_i}\right)\right)\\
&\overset{\text{(a)}}{\geq} (\mu_i-\nu_i)\left(2\sqrt{2c}-5/2-8\log\frac{n}{\beta}\right)
\end{aligned}$$

Here, step (a) is due to the fact that $f(x) = x\log\frac{n}{\beta x}$ is monotone increasing for $n \geq 3$. When $n < 3$, we have $\mu_i = 1$ and $\nu_i = 0$ (otherwise, there is no uncertainty left in Y) and hence the problem becomes trivial.

- $1/n \leq \nu_i \leq 1/2 \leq \mu_i$.
  In this case, we cannot replace p_i, q_i with μ_i or ν_i. However, notice that $\max\{\mu_i(1-\mu_i),\nu_i(1-\nu_i)\} \leq 1/4$, we have

$$\begin{aligned}
\mathrm{LB2}_i &\geq \mu_i(1-\mu_i)+\nu_i(1-\nu_i) - 5(\mu_i-\nu_i)^2\\
&\quad - 8(\mu_i-\nu_i)\left(\nu_i\log\frac{n}{\beta}+\nu_i\log\frac{1}{\nu_i}\right) + 8c\,(\mu_i-\nu_i)^2\\
&= \mu_i(1-\mu_i)+\nu_i(1-\nu_i)+(\mu_i-\nu_i)^2+(8c-6)(\mu_i-\nu_i)^2\\
&\quad - 8(\mu_i-\nu_i)\left(\nu_i\log\frac{n}{\beta}+\nu_i\log\frac{1}{\nu_i}\right)\\
&= \mu_i(1-\nu_i)+\nu_i(1-\mu_i)+(8c-6)(\mu_i-\nu_i)^2\\
&\quad - 8(\mu_i-\nu_i)\left(\nu_i\log\frac{n}{\beta}+\nu_i\log\frac{1}{\nu_i}\right)\\
&\geq \mu_i(1-\nu_i)+(8c-6)(\mu_i-\nu_i)^2 - 8(\mu_i-\nu_i)\left(\nu_i\log\frac{n}{\beta}+\nu_i\log\frac{1}{\nu_i}\right) \qquad \text{(B.16)}\\
&\overset{\nu_i\geq 1/n}{\geq} \mu_i(1-\nu_i)+(8c-6)(\mu_i-\nu_i)^2 - 8(\mu_i-\nu_i)\nu_i\log\frac{n^2}{\beta}
\end{aligned}$$

To further simplify notation, we denote $\gamma_1 \triangleq 8c-6$, and $\gamma_2 \triangleq 8\log\frac{n^2}{\beta}$. Then the above equation can be rewritten as

$$\mathrm{LB2}_i \geq \mu_i(1-\nu_i)+\gamma_1(\mu_i-\nu_i)^2 - \gamma_2(\mu_i-\nu_i)\nu_i$$

If $\mu_i - \nu_i \le \frac{1}{2\gamma_2}$, then

$$\mathrm{LB2}_i \ge \mu_i(1-\nu_i) + \gamma_1(\mu_i-\nu_i)^2 - \frac{1}{2\gamma_2}\gamma_2\nu_i = \mu_i(1-\nu_i) - \frac{\nu_i}{2} \ge 0$$

Otherwise, if $\mu_i - \nu_i > \frac{1}{2\gamma_2}$, we have

$$\begin{aligned}\mathrm{LB2}_i &\ge \mu_i(1-\nu_i) + (\mu_i-\nu_i)\left(\gamma_1(\mu_i-\nu_i) - \gamma_2\nu_i\right)\\ &> \mu_i(1-\nu_i) + (\mu_i-\nu_i)\left(\gamma_1\frac{1}{2\gamma_2} - \gamma_2\nu_i\right)\\ &> \frac{\mu_i-\nu_i}{2}\left(\frac{\gamma_1}{\gamma_2} - \gamma_2\right)\end{aligned}$$

- $\nu_i \le 1/n < 1/2 \le \mu_i$. In this case, we have

$$\begin{aligned}\mathrm{LB2}_i &\overset{\text{Eq (B.16)}}{\ge} \mu_i(1-\nu_i) + \gamma_1(\mu_i-\nu_i)^2 - 8(\mu_i-\nu_i)\left(\nu_i\log\frac{n}{\beta} + \nu_i\log\frac{1}{\nu_i}\right)\\ &\ge \mu_i(1-\nu_i) + \gamma_1(\mu_i-\nu_i)^2 - 8(\mu_i-\nu_i)\left(\frac{1}{n}\log\frac{n}{\beta} + \frac{\log n}{n}\right)\\ &= \mu_i(1-\nu_i) + \gamma_1(\mu_i-\nu_i)^2 - \frac{\gamma_2}{n}(\mu_i-\nu_i)\\ &> \mu_i(1-\nu_i) + (\mu_i-\nu_i)\left(\gamma_1\frac{n-2}{2n} - \frac{\gamma_2}{n}\right)\\ &\overset{\text{(a)}}{\ge} \frac{\mu_i-\nu_i}{3}\left(\frac{\gamma_1}{2} - \gamma_2\right)\\ &\ge \frac{\mu_i-\nu_i}{3}\left(\frac{\gamma_1}{\gamma_2} - \gamma_2\right)\end{aligned}$$

  Step (a) is due to the fact that $1/n < 1/2$ and therefore $n \ge 3$.

Putting the above cases together, we obtain the following equations:

$$\mathrm{LB2}_i \ge \begin{cases} \frac{(\mu_i-\nu_i)\left(2\sqrt{2c} - \frac{5}{2} - 8\log\frac{n}{\beta}\right)}{} & \text{if } \mu_i \ge \nu_i \ge 1/2, \text{or } \nu_i \le \mu_i \le 1/2\\ 0 & \text{if } \frac{1}{n} \le \nu_i \le \frac{1}{2} \le \mu_i, \text{ and } \mu_i - \nu_i \le \frac{1}{2\gamma_2}\\ \frac{\mu_i-\nu_i}{2}\left(\frac{\gamma_1}{\gamma_2} - \gamma_2\right) & \text{if } \frac{1}{n} \le \nu_i \le \frac{1}{2} \le \mu_i, \text{ and } \mu_i - \nu_i > \frac{1}{2\gamma_2}\\ \frac{\mu_i-\nu_i}{3}\left(\frac{\gamma_1}{\gamma_2} - \gamma_2\right) & \text{if } \nu_i \le \frac{1}{n} < \frac{1}{2} \le \mu_i \end{cases}$$

Fix $\eta \ge 0$. Let $c = 8\left(\log\frac{2n^2}{\eta}\right)^2$, we have $\gamma_1 > \left(8\log\frac{n^2}{\eta}\right)^2$, and $\gamma_2 = 8\log\frac{n^2}{\beta}$, so

$$\frac{\gamma_1}{\gamma_2} - \gamma_2 = \frac{(\sqrt{\gamma_1}-\gamma_2)(\sqrt{\gamma_1}+\gamma_2)}{\gamma_2} > 8\frac{\sqrt{\gamma_1}+\gamma_2}{\gamma_2}\log\frac{\beta}{\eta}$$

and thus we get

$$\mathrm{LB2}_i \ge \begin{cases} 8(\mu_i-\nu_i)\log\frac{\beta}{\eta} & \text{if } \mu_i \ge \nu_i \ge 1/2, \text{ or } \nu_i \le \mu_i \le 1/2\\ 0 & \text{if } 1/n \le \nu_i \le 1/2 \le \mu_i, \text{ and } \mu_i - \nu_i \le \frac{1}{2\gamma_2}\\ \frac{4(\mu_i-\nu_i)(\sqrt{\gamma_1}+\gamma_2)}{\gamma_2}\log\frac{\beta}{\eta} & \text{if } \nu_i \le 1/2 \le \mu_i, \text{ and } \mu_i - \nu_i > \frac{1}{2\gamma_2} \end{cases}$$

That is, if $\beta \geq \eta$, we have $\mathrm{LB2}_i \geq 0$ for all $i \in \{1, \ldots, t\}$.

On the other hand, if $\beta < \eta$, we get $\frac{4(\sqrt{\gamma_1}+\gamma_2)}{\gamma_2} = \frac{4(\log\frac{n^2}{\eta}+\log\frac{n^2}{\beta})}{\log\frac{n^2}{\beta}} \leq 8$, and therefore $\mathrm{LB2}_i \geq 8(\mu_i - \nu_i)\log\frac{\beta}{\eta}$.

Summing over all $i \in \{1, \ldots, t\}$, we get that for $\beta < \eta$, it holds $\mathrm{LB2} \geq \sum_{y_i \in \mathcal{Y}} |\mu_i - \nu_i| \cdot \frac{2\alpha^2\beta^2(1-2\epsilon)^2}{h_+ h_-} \log\frac{\beta}{\eta}$. We hence get

$$\mathrm{LB2} \geq \begin{cases} -2t\,(1-2\epsilon)^2\,\alpha\beta\log\frac{\eta}{\alpha\beta} & \text{if } \alpha\beta < \eta \\ 0 & \text{if } \alpha\beta \geq \eta \end{cases}$$

Further relaxing the above condition by $\alpha\beta\log\frac{\eta}{\alpha\beta} \leq \eta - \alpha\beta \leq \eta$, we obtain:

$$\mathrm{LB2} \geq -2t\,(1-2\epsilon)^2\,\eta \tag{B.17}$$

Combining Equation (B.11), (B.14), and (B.17), we get

$$\Delta_{\mathrm{AUX}} \geq \frac{1}{16}\,(1-2\epsilon)^2\,\Delta_{\mathrm{EC}^2} - 2t\,(1-2\epsilon)^2\,\eta. \tag{B.18}$$

Hence, we have related $\Delta_{\mathrm{AUX}}(v \mid \psi)$ to $\Delta_{\mathrm{EC}^2,\psi}(v)$, as stated in Lemma 3.

B.3.6. Bounding Δ_{AUX} *against* Δ_{ECED}

To finish the proof for Lemma 3, it remains to bound Δ_{AUX} against Δ_{ECED}. In this subsection, we complete the proof of Lemma 3, by showing that $\Delta_{\mathrm{AUX}}(X_e \mid \psi) + 2t\,(1-2\epsilon)^2\,\eta \geq \Delta_{\mathrm{ECED},\psi}(X_e)/64$.

Recall that ϵ is the noise rate of test e. Let $\rho = \frac{\epsilon}{1-\epsilon}$ be the discount factor for inconsistent root-causes. By the definition of Δ_{ECED} in Equation (3.1), we first expand the expected offset value of performing test e:

$$\mathbb{E}_{x_e}[\delta_{\mathrm{OFFSET}}(x_e)] = \sum_{y_i \in \mathcal{Y}} (\alpha_i + \beta_i)(1 - \alpha_i - \beta_i)\epsilon\left(1 - \rho^2\right).$$

Denote $\gamma = \epsilon\left(1-\rho^2\right)$. Then, we can expand Δ_{ECED} as

$$\Delta_{\mathrm{ECED}} = \sum_{y_i \in \mathcal{Y}} \Bigg(\overbrace{(\alpha_i + \beta_i)(1 - \alpha_i - \beta_i)\,(1-\gamma)}^{\text{(initial total edge weight)}-\text{(offset value)}} - \overbrace{(h_+(\alpha_i + \rho\beta_i)(\alpha + \rho\beta - \alpha_i - \rho\beta_i) + h_-(\beta_i + \rho\alpha_i)(\beta + \rho\alpha - \beta_i - \rho\alpha_i))}^{\text{expected remaining weight after discounting}} \Bigg)$$

$$= h_+ \sum_{y_i \in \mathcal{Y}} (-\gamma\alpha_i(\alpha - \alpha_i) + \alpha_i(\beta - \beta_i)(1 - \gamma - \rho) + \beta_i(\alpha - \alpha_i)(1 - \gamma - \rho) + \beta_i(\beta - \beta_i)(1 - \gamma - \rho^2)) +$$

$$h_- \sum_{y_i \in \mathcal{Y}} (-\gamma\beta_i(\beta-\beta_i) + \beta_i(\alpha-\alpha_i)(1-\gamma-\rho) + \alpha_i(\beta-\beta_i)(1-\gamma-\rho) + \alpha_i(\alpha-\alpha_i)(1-\gamma-\rho^2))$$
$$= \sum_{y_i \in \mathcal{Y}} \left(2(1-\gamma-\rho)\alpha_i(\beta-\beta_i) + \left(h_+(1-\gamma-\rho^2) - h_-\gamma\right)\beta_i(\beta-\beta_i) + \left(h_-(1-\gamma-\rho^2) - h_+\gamma\right)\alpha_i(\alpha-\alpha_i)\right)$$

Since $\gamma = \frac{\epsilon(1-2\epsilon)}{(1-\epsilon)^2}$, $1-\gamma-\rho^2 = \frac{1-2\epsilon}{1-\epsilon}$, and $1-\gamma-\rho = \left(\frac{1-2\epsilon}{1-\epsilon}\right)^2$, we have,

$$h_+(1-\gamma-\rho^2) - h_-\gamma = (\alpha(1-\epsilon)+\beta\epsilon)\frac{1-2\epsilon}{1-\epsilon} - (\alpha\epsilon+\beta(1-\epsilon))\frac{\epsilon(1-2\epsilon)}{(1-\epsilon)^2}$$
$$= \left(\frac{1-2\epsilon}{1-\epsilon}\right)^2 \alpha$$

Therefore

$$\Delta_{\text{ECED}}$$
$$= \left(\frac{1-2\epsilon}{1-\epsilon}\right)^2 \left(\alpha \sum_{y_i \in \mathcal{Y}} \beta_i(\beta-\beta_i) + \beta \sum_{y_i \in \mathcal{Y}} \alpha_i(\alpha-\alpha_i) + 2\sum_{y_i \in \mathcal{Y}} \alpha_i(\beta-\beta_i)\right)$$
$$= \left(\frac{1-2\epsilon}{1-\epsilon}\right)^2 \Delta_{\text{EC}^2} \tag{B.19}$$

Combining Equation (B.19) with Equation (B.18) we obtain

$$\Delta_{\text{AUX}} + 2t\,(1-2\epsilon)^2\,\eta \geq \frac{(1-\epsilon)^2}{16}\Delta_{\text{ECED}}$$
$$= \frac{1}{16}\,(1-2\epsilon)^2\,\Delta_{\text{EC}^2}$$

With the results from Appendix §B.3.5 and §B.3.6, we therefore complete the proof of Lemma 3.

B.4. Proof of Theorem 1 part 3: Relating ECED *to* OPT

B.4.1. Bounding the error probability: Noiseless vs. noisy setting

Now that we have seen how ECED interacts with our auxiliary function in terms of the one-step gain, it remains to understand how one can relate the one-step gain to the gain of an optimal policy $\Delta_{\text{AUX}}(\text{OPT} \mid \psi)$, over k steps. In this subsection, we make an important step towards this goal.

Specifically, we provide

Lemma 7. *Consider a policy π of length k, and assume that we are using a stochastic estimator (*SE*). Let $p_{\mathrm{E}}^{\top}$ be the error probability of* SE *before running policy π, $p_{\mathrm{E},noisy}^{\perp}$ be the average error probability of* SE *after running π in the noisy setting, and $p_{\mathrm{E},noiseless}^{\perp}$ be the average error probability of* SE *after running π in the noiseless setting. Then*

$$p_{\mathrm{E},noiseless}^{\perp} \leq p_{\mathrm{E},noisy}^{\perp}$$

Proof of Lemma 7. Recall that a *stochastic estimator* predicts the value of a random variable, by randomly drawing from its distribution. Let π be a policy. We denote by $p_{\mathrm{E}}(\pi_\phi)$ the expected error probability of an stochastic estimator after observing π_ϕ:

$$p_{\mathrm{E,noisy}}^{\perp} = \mathbb{E}_{\phi}[p_{\mathrm{E}}(\pi_\phi)] = \sum_{\phi} p(\pi_\phi) \sum_{y \in \mathcal{Y}} p(y \mid \pi_\phi)(1 - p(y \mid \pi_\phi))$$

where $\phi \in \mathcal{V} \times \mathcal{O}$ denotes a set of test-outcome pairs, and π_ϕ denotes a path taken by π, given that it observes ϕ.

Now, let us see what happens in the noiseless setting: we run π exactly as it is, but in the end compute the error probability of the noiseless setting (i.e., as if we know which test outcomes are corrupted by noise). Denote the noise put on the tests by Ξ, and the realized noise by ξ. We can imagine the noiseless setting through the following equivalent way: we ran the same policy π exactly as in the noisy setting. But upon completion of π we reveal what Ξ was. We thus have

$$p(y \mid \pi_\phi) = \sum_{\Xi = \xi} p(y \mid \pi_\phi, \xi) p(\xi \mid \pi)$$

The error probability upon observing π_ϕ and $\Xi = \xi$ is

$$p_{\mathrm{E}}(\pi_\phi, \xi) = \sum_{y \in \mathcal{Y}} p(y \mid \pi_\phi, \xi)(1 - p(y \mid \pi_\phi, \xi)).$$

The expected error probability in the noiseless setting after running π is

$$p_{\mathrm{E,noiseless}}^{\perp} = \mathbb{E}_{\phi,n}[p_{\mathrm{E}}(\pi_\phi, \xi)] = \sum_{\phi,n} p(\pi_\phi, \xi) \sum_{y \in \mathcal{Y}} p(y \mid \pi_\phi, \xi)(1 - p(y \mid \pi_\phi, \xi)) \tag{B.20}$$

Now, we can relate $p_{\mathrm{E,noisy}}^{\perp}$ to $p_{\mathrm{E,noiseless}}^{\perp}$.

$$\begin{aligned}
p_{\mathrm{E,noisy}}^{\perp} &= \sum_{\phi} p(\pi_\phi) \sum_{y \in \mathcal{Y}} p(y \mid \pi_\phi)(1 - p(y \mid \pi_\phi)) \\
&= \sum_{\phi} p(\pi_\phi) \sum_{y \in \mathcal{Y}} \sum_{\xi} p(\xi \mid \pi_\phi) p(y \mid \pi_\phi, \xi)(1 - \sum_{n} p(\xi \mid \pi_\phi) p(y \mid \pi_\phi, \xi)) \\
&\overset{(a)}{\geq} \sum_{\phi} p(\pi_\phi) \sum_{y \in \mathcal{Y}} \sum_{\xi} p(\xi \mid \pi_\phi) p(y \mid \pi_\phi, \xi)(1 - p(y \mid \pi_\phi, \xi)) \\
&= \sum_{\phi,\xi} p(\pi_\phi, \xi) \sum_{y \in \mathcal{Y}} p(y \mid \pi_\phi, \xi)(1 - p(y \mid \pi_\phi, \xi))
\end{aligned}$$

where (a) is by Jensen's inequality and the fact that $f(x) = x(1-x)$ is concave. Combining with Equation (B.20) we complete the proof. □

Essentially, Lemma 7 implies that, in terms of the reduction in the expected prediction error of SE, running a policy in the noise-free setting has higher gain than running the exact same policy in the noisy setting. This result is important to us, since analyzing a policy in the noise-free setting is often easier. We are going to use Lemma 7 in the next section, to relate the gain of an optimal policy $\Delta_{\mathrm{EC}^2,\psi}(\mathrm{OPT})$ in the EC^2 objective (which assumes tests to be noise-free), with the gain $\Delta_{\mathrm{AUX}}(\mathrm{OPT} \mid \psi)$ in the auxiliary function (which considers noisy test outcomes).

B.4.2. Key lemma: One-step gain of ECED *VS. k-step gain of* OPT

Now we are ready to state our key lemma, which connects $\Delta_{\mathrm{AUX}}(v \mid \psi)$ to $\Delta_{\mathrm{AUX}}(\mathrm{OPT} \mid \psi)$.

Lemma 8 (Key Lemma)**.** *Fix $\eta, \tau \in (0,1)$. Let $n = |\mathrm{supp}(\Theta)|$ be the number of root-causes, $t = |\mathcal{Y}|$ be the number of target values, $\mathsf{OPT}(\delta_{\mathsf{opt}})$ be the optimal policy that achieves $p_{\mathrm{ERR}}(\mathsf{OPT}(\delta_{\mathsf{opt}})) \le \delta_{\mathsf{opt}}$, and ψ_ℓ be the partial realization observed by running* ECED *with cost ℓ. We denote by $f^{avg}_{\mathrm{AUX}}(\ell) := \mathbb{E}_{\psi_\ell}[f_{\mathrm{AUX}}(\psi_\ell)]$ the expected value of $f_{\mathrm{AUX}}(\psi_\ell)$ over all the paths ψ_ℓ at cost ℓ. Assume that $f^{avg}_{\mathrm{AUX}}(\ell) \le \delta_{\mathsf{g}}$. We then have*

$$f^{avg}_{\mathrm{AUX}}(\ell) - f^{avg}_{\mathrm{AUX}}(\ell+1) \ge \frac{f^{avg}_{\mathrm{AUX}}(\ell) - \delta_{\mathsf{opt}}}{k} \cdot \frac{c_\epsilon}{c_\delta} + c_{\eta,\epsilon}.$$

where $k = \mathrm{cost}(\mathsf{OPT}(\delta_{\mathsf{opt}})))$, $c_{\eta,\epsilon} \triangleq 2t(1-2\epsilon)^2\eta$, $c_\delta \triangleq (6c+8)\log(n/\delta_{\mathsf{g}})$, $c \triangleq 8\left(\log(2n^2/\eta)\right)^2$, and $c_\epsilon \triangleq (1-2\epsilon)^2/16$.

Proof of Lemma 8. Let ψ_ℓ be a path ending up at level ℓ of the greedy algorithm. Recall that $\Delta_{\mathrm{EC}^2}(X_e \mid \psi_\ell)$ denotes the gain in f_{EC^2} if we perform test e and assuming it to be *noiseless* (i.e., we perform edge cutting as if the outcome of test e is noiseless), conditioning on partial observation ψ_ℓ. Further, recall that $\Delta_{\mathrm{AUX}}(X_e \mid \psi_\ell)$ denotes the gain in f_{AUX} if we perform *noisy* test e after observing ψ_ℓ and perform Bayesian update on the root-causes.

Let $e = \arg\max_{e'} \Delta_{\mathrm{ECED}}(X_{e'} \mid \psi_\ell)$ be the test chosen by ECED, and $\hat{e} = \arg\max_{e'} \Delta_{\mathrm{EC}^2}(X_{e'} \mid \psi_\ell)$ be the test that maximizes Δ_{EC^2}, then by Lemma 3 we know

$$\begin{aligned}
\Delta_{\mathrm{AUX}}(X_e \mid \psi_\ell) + c_{\eta,\epsilon} &\ge \frac{(1-\epsilon)^2}{16}\left(\Delta_{\mathrm{ECED},\psi_\ell}(X_e)\right) \\
&\ge \frac{(1-\epsilon)^2}{16}\left(\Delta_{\mathrm{ECED},\psi_\ell}(X_{\hat{e}})\right) \\
&= \frac{1}{16}(1-2\epsilon)^2\,\Delta_{\mathrm{EC}^2,\psi}(X_{\hat{e}}) \qquad \text{(B.21)}
\end{aligned}$$

Note that $\Delta_{\mathrm{EC}^2,\psi_\ell}(X_e)$ is the EC2 gain of test e over the *normalized* edge weights at step $\ell+1$ in the noiseless setting. That is, upon observing ψ_ℓ, we create a new EC2 problem instance (by considering the posterior probability

over root-causes at ψ_ℓ), and run (noiseless) greedy algorithm w.r.t. the EC2 objective on such problem instance. Recall that $c_\epsilon \triangleq (1-2\epsilon)/16$. By adaptive submodularity of f_{EC^2} (in the noiseless setting, see Golovin et al. [14]), we obtain

$$\max_e \Delta_{\mathrm{EC}^2,\psi}(X_e) \overset{\substack{\text{adaptive}\\ \text{submodularity}}}{\geq} \frac{f^{\top}_{\mathrm{EC}^2,\psi_\ell} - \mathbb{E}[f^{\perp}_{\mathrm{EC}^2,\psi_\ell}]}{k}$$

where by $f^{\top}_{\mathrm{EC}^2,\psi_\ell}$ we mean the initial EC2 objective value given partial realization ψ_ℓ, and by $\mathbb{E}[f^{\perp}_{\mathrm{EC}^2,\psi_\ell}]$ we mean the expected gain in f_{EC^2} when we run $\mathsf{OPT}\,(\delta_{\mathsf{opt}})$. Note that $\mathsf{OPT}\,(\delta_{\mathsf{opt}})$ has worst-case length k.

Now, imagine that we run the policy $\mathsf{OPT}\,(\delta_{\mathsf{opt}})$, and upon completion of the policy we can observe the noise. We consider the gain of such policy in f_{EC^2}:

$$f^{\top}_{\mathrm{EC}^2} - \mathbb{E}[f^{\perp}_{\mathrm{EC}^2}] \overset{\text{(a)}}{=} p^{\top}_{\mathrm{E}} - \mathbb{E}[f^{\perp}_{\mathrm{EC}^2}] \overset{\text{(b)}}{\geq} p^{\top}_{\mathrm{E}} - p^{\perp}_{\mathrm{E,noiseless}}.$$

The reason for step (a) is that the error probability of the stochastic estimator upon observing ψ_ℓ, i.e., $p^{\top}_{\mathrm{E}}$, is equivalent to the total amount of edge weight at ψ_ℓ, i.e., $f^{\top}_{\mathrm{EC}^2,\psi_\ell}$. The reason for step (b) is that under the noiseless setting (i.e., assuming we have access to the noise), the EC2 objective is always a lower-bound on the error probability of the *stochastic estimator* (due to normalization). Thus, $\mathbb{E}[f^{\perp}_{\mathrm{EC}^2}] \leq p^{\perp}_{\mathrm{E,noiseless}}$.

Hence we get

$$\Delta_{\mathrm{AUX}}(X_e \mid \psi) + c_{\eta,\epsilon} \geq c_\epsilon \frac{p^{\top}_{\mathrm{E},\psi_\ell} - p^{\perp}_{\mathrm{E,noiseless},\psi_\ell}}{k}.$$

Here $p^{\top}_{\mathrm{E},\psi_\ell}$ denotes the error probability under $\mathbb{P}\left[Y \mid \psi_\ell\right]$, and $p^{\perp}_{\mathrm{E,noisy},\psi_\ell}$ denotes the expected error probability of running $\mathsf{OPT}\,(\delta_{\mathsf{opt}})$ after ψ_ℓ in the *noise-free* setting. By Lemma 7 we get

$$\Delta_{\mathrm{AUX}}(X_e \mid \psi) + c_{\eta,\epsilon} \geq c_\epsilon \frac{p^{\top}_{\mathrm{E},\psi_\ell} - p^{\perp}_{\mathrm{E,noisy},\psi_\ell}}{k},$$

where $p^{\perp}_{\mathrm{E,noisy},\psi_\ell}$ denotes the expected error probability of running $\mathsf{OPT}\,(\delta_{\mathsf{opt}})$ after ψ_ℓ in the *noisy* setting. By (the lower bound in) Lemma 4, we know that $p^{\top}_{\mathrm{E},\psi_\ell} = p_{\mathrm{E}}(\psi_\ell) \geq p^{\mathrm{MAP}}_{\mathrm{ERR}}(\psi_\ell)$, and hence

$$\Delta_{\mathrm{AUX}}(X_e \mid \psi) + c_{\eta,\epsilon} \geq c_\epsilon \frac{p^{\mathrm{MAP}}_{\mathrm{ERR}}(\psi_\ell) - \delta_{\mathsf{opt}}}{k},$$

Taking expectation with respect to ψ_ℓ, we get

$$\mathbb{E}_{\psi_\ell}[\Delta_{\mathrm{AUX}}(X_e \mid \psi) + c_{\eta,\epsilon}] \geq c_\epsilon \frac{\mathbb{E}_{\psi_\ell}\left[p^{\mathrm{MAP}}_{\mathrm{ERR}}(\psi_\ell)\right] - \delta_{\mathsf{opt}}}{k}. \tag{B.22}$$

Using (the upper bound in) Lemma 2, we obtain

$$
\begin{aligned}
f_{\text{AUX}}^{\text{avg}}(\ell) &= \mathbb{E}_{\psi_\ell}[f_{\text{AUX}}(\psi_\ell)] \\
&\le (3c+4)\left(\mathbb{E}_{\psi_\ell}\left[\mathbb{H}_2\left(p_{\text{ERR}}^{\text{MAP}}(\psi_\ell)\right)\right] + \mathbb{E}_{\psi_\ell}\left[p_{\text{ERR}}^{\text{MAP}}(\psi_\ell)\right]\log n\right) \\
&\overset{\text{(a)}}{\le} (3c+4)\left(\mathbb{H}_2\left(\mathbb{E}_{\psi_\ell}\left[p_{\text{ERR}}^{\text{MAP}}(\psi_\ell)\right]\right) + \mathbb{E}_{\psi_\ell}\left[p_{\text{ERR}}^{\text{MAP}}(\psi_\ell)\right]\log n\right) \qquad \text{(B.23)}
\end{aligned}
$$

where (a) is by Jensen's inequality.

Suppose we run ECED, and achieve expected error probability δ_{g}, then clearly before ECED terminates we have $\mathbb{E}_{\psi_\ell}\left[p_{\text{ERR}}^{\text{MAP}}(\psi_\ell)\right] \ge \delta_{\mathsf{g}}$.

Assuming $\mathbb{E}_{\psi_\ell}\left[p_{\text{ERR}}^{\text{MAP}}(\psi_\ell)\right] \le 1/2$, we have

$$
\begin{aligned}
f_{\text{AUX}}^{\text{avg}}(\ell) &\le (3c+4)\mathbb{E}_{\psi_\ell}\left[p_{\text{ERR}}^{\text{MAP}}(\psi_\ell)\right]\left(2\log\frac{1}{\mathbb{E}_{\psi_\ell}[p_{\text{ERR}}^{\text{MAP}}(\psi_\ell)]} + \log n\right) \\
&\le (3c+4)\mathbb{E}_{\psi_\ell}\left[p_{\text{ERR}}^{\text{MAP}}(\psi_\ell)\right]\left(2\log\frac{1}{\delta_{\mathsf{g}}} + \log n\right) \\
&\le \mathbb{E}_{\psi_\ell}\left[p_{\text{ERR}}^{\text{MAP}}(\psi_\ell)\right]\cdot(6c+8)\log\frac{n}{\delta_{\mathsf{g}}} \qquad \text{(B.24)}
\end{aligned}
$$

which gives us

$$
\mathbb{E}_{\psi_\ell}\left[p_{\text{ERR}}^{\text{MAP}}(\psi_\ell)\right] \ge \frac{f_{\text{AUX}}^{\text{avg}}(\ell)}{(6c+8)\log\frac{n}{\delta_{\mathsf{g}}}} \overset{c_\delta \triangleq (6c+8)\log\frac{n}{\delta_{\mathsf{g}}}}{=} \frac{f_{\text{AUX}}^{\text{avg}}(\ell)}{c_\delta}. \qquad \text{(B.25)}
$$

Combining Equation (B.25) with Equation (B.22), we get

$$
\begin{aligned}
f_{\text{AUX}}^{\text{avg}}(\ell) - f_{\text{AUX}}^{\text{avg}}(\ell+1) &= \mathbb{E}_{\psi_\ell}[\Delta_{\text{AUX}}(e \mid \psi)] \\
&\ge c_\epsilon \frac{\frac{f_{\text{AUX}}^{\text{avg}}(\ell)}{c_\delta} - \delta_{\mathsf{opt}}}{k} - c_{\eta,\epsilon} \\
&= \frac{f_{\text{AUX}}^{\text{avg}}(\ell) - \delta_{\mathsf{opt}} c_\delta}{k}\cdot\frac{c_\epsilon}{c_\delta} - c_{\eta,\epsilon}
\end{aligned}
$$

which completes the proof. □

B.5. *Proof of Theorem 1 final step: Near-optimality of* ECED

We are going to put together the pieces from the previous subsections, to give a proof of our main theoretical result (Theorem 1).

Proof of Theorem 1. In the following, we use both $\mathsf{OPT}_{[k]}$ and $\mathsf{OPT}(\delta_{\mathsf{opt}})$ to represent the optimal policy that achieves prediction error δ_{opt}, with worst-cast cost (i.e., length) k. Define $S(\pi,\phi)$ to be the (partial) realization seen by policy π under realization ϕ. With a slight abuse of notation, we use $f_{\text{AUX}}^{\text{avg}}\left(\mathsf{OPT}_{[k]}\right) := \mathbb{E}_\phi\left[f_{\text{AUX}}(S(\mathsf{OPT}_{[k]},\phi))\right]$ to denote the expected value achieved by running $\mathsf{OPT}_{[k]}$.

After running $\mathsf{OPT}_{[k]}$, we know by Lemma 2 that the expected value of f_{AUX} is lower bounded by $2c\cdot\delta_{\mathsf{opt}}$. That is, $\delta_{\mathsf{opt}}\cdot c_\delta \le f_{\text{AUX}}^{\text{avg}}\left(\mathsf{OPT}_{[k]}\right)\cdot\frac{c_\delta}{2c} \le f_{\text{AUX}}^{\text{avg}}\left(\mathsf{OPT}_{[k]}\right)\cdot$

$4\log(n/\delta_{\mathsf{g}})$, where the last inequality is due to $c_\delta \triangleq (6c+8)\log\frac{n}{\delta_{\mathsf{g}}} < 8c\log\frac{n}{\delta_{\mathsf{g}}}$. We then have

$$\begin{aligned} f^{\mathrm{avg}}_{\mathrm{AUX}}(\ell) - f^{\mathrm{avg}}_{\mathrm{AUX}}(\ell+1) &\overset{\text{Lemma 8}}{\geq} (f^{\mathrm{avg}}_{\mathrm{AUX}}(\ell) - \delta_{\mathsf{opt}}\cdot c_\delta)\cdot\frac{c_\varepsilon}{kc_\delta} - c_{\eta,\epsilon} \\ &\geq \left(f^{\mathrm{avg}}_{\mathrm{AUX}}(\ell) - f^{\mathrm{avg}}_{\mathrm{AUX}}\left(\mathsf{OPT}_{[k]}\right)\cdot 4\log\frac{n}{\delta_{\mathsf{g}}}\right)\cdot\frac{c_\varepsilon}{kc_\delta} - c_{\eta,\epsilon} \quad \text{(B.26)} \end{aligned}$$

Let $\Delta_\ell \triangleq f^{\mathrm{avg}}_{\mathrm{AUX}}(\ell) - f^{\mathrm{avg}}_{\mathrm{AUX}}\left(\mathsf{OPT}_{[k]}\right)\cdot 4\log\frac{n}{\delta_{\mathsf{g}}}$, so that Inequality (B.26) implies $\Delta_\ell - \Delta_{\ell+1} \geq \Delta_\ell\cdot\frac{c_\epsilon}{kc_\delta} - c_{\eta,\epsilon}$. From here we get $\Delta_{\ell+1} \leq \left(1-\frac{c_\epsilon}{kc_\delta}\right)\Delta_\ell + c_{\eta,\epsilon}$, and hence

$$\begin{aligned} \Delta_{k'} &\leq \left(1-\frac{c_\epsilon}{kc_\delta}\right)^{k'}\Delta_0 + \sum_{i=0}^{k'}\left(1-\frac{c_\epsilon}{kc_\delta}\right)^{i}\cdot c_{\eta,\epsilon} \\ &\overset{(a)}{\leq} \exp\left(-k'\frac{c_\epsilon}{kc_\delta}\right)\Delta_0 + \frac{1-\left(1-\frac{c_\epsilon}{kc_\delta}\right)^{k'}}{\frac{c_\epsilon}{kc_\delta}}\cdot c_{\eta,\epsilon} \\ &\overset{(b)}{\leq} \exp\left(-k'\frac{c_\epsilon}{kc_\delta}\right)\Delta_0 + \frac{kc_\delta}{c_\epsilon}\cdot c_{\eta,\epsilon} \end{aligned}$$

where step (a) is due to the fact that $(1-x)^{k'} \leq \exp(-k'x)$ for any $x<1$, and step (b) is due to $\left(1-\frac{c_\epsilon}{kc_\delta}\right)^{k'} > 0$. It follows that

$$\begin{aligned} &f^{\mathrm{avg}}_{\mathrm{AUX}}(k') - f^{\mathrm{avg}}_{\mathrm{AUX}}\left(\mathsf{OPT}_{[k]}\right)\cdot 4\log\frac{n}{\delta_{\mathsf{g}}} \\ &\leq \exp\left(-k'\frac{c_\epsilon}{kc_\delta}\right)\Delta_0 + \frac{kc_\delta}{c_\epsilon}\cdot c_{\eta,\epsilon} \\ &\leq \exp\left(-k'\frac{c_\epsilon}{kc_\delta}\right)\left(f^{\mathrm{avg}}_{\mathrm{AUX}}(\emptyset) - f^{\mathrm{avg}}_{\mathrm{AUX}}\left(\mathsf{OPT}_{[k]}\right)\cdot 4\log\frac{n}{\delta_{\mathsf{g}}}\right) + \frac{kc_\delta}{c_\epsilon}\cdot c_{\eta,\epsilon} \end{aligned}$$

This gives us

$$\begin{aligned} f^{\mathrm{avg}}_{\mathrm{AUX}}(k') \leq\ & \underbrace{f^{\mathrm{avg}}_{\mathrm{AUX}}(\emptyset)\cdot\exp\left(-k'\frac{c_\epsilon}{kc_\delta}\right)}_{\mathrm{UB1}} \\ &+ \underbrace{f^{\mathrm{avg}}_{\mathrm{AUX}}\left(\mathsf{OPT}_{[k]}\right)\cdot 4\log\frac{n}{\delta_{\mathsf{g}}}\left(1-\exp\left(-k'\frac{c_\epsilon}{kc_\delta}\right)\right)}_{\mathrm{UB2}} \\ &+ \underbrace{\frac{kc_\delta}{c_\epsilon}\cdot c_{\eta,\epsilon}}_{\mathrm{UB3}} \end{aligned} \quad \text{(B.27)}$$

Denote the three terms on the RHS. of Equation (B.27) as UB1, UB2 and UB3, respectively. We get

$$\begin{cases} \text{UB1} & \overset{\text{Eq (B.23)}}{\leq} (3c+4)\,(1+\log n)\cdot \exp\left(-k'\frac{c_\epsilon}{kc_\delta}\right) \\ \text{UB2} & \overset{\text{Eq (B.24)}}{<} (6c+8)\cdot \delta_{\mathsf{opt}} \log\frac{n}{\delta_{\mathsf{opt}}}\cdot 4\log\frac{n}{\delta_{\mathsf{g}}} \\ \text{UB3} & = k\cdot(6c+8)\log\frac{n}{\delta_{\mathsf{g}}}\cdot\frac{2t(1-2\epsilon)^2\eta}{\frac{1}{16}(1-2\epsilon)^2} = (6c+8)\cdot 32\cdot k\cdot\log\frac{n}{\delta_{\mathsf{g}}}\cdot t\eta \end{cases}$$

Now we set

$$\begin{cases} k' & \triangleq \frac{kc_\delta}{c_\varepsilon}\cdot\ln\frac{8\log n}{\delta_{\mathsf{g}}} \\ \delta_{\mathsf{opt}} & \triangleq \frac{\delta_{\mathsf{g}}}{64\cdot 36\cdot\log n\cdot\log\frac{1}{\delta_{\mathsf{g}}}\cdot\log\frac{n}{\delta_{\mathsf{g}}}} \end{cases} \tag{B.28}$$

and obtain $\exp\left(-k'\frac{c_\epsilon}{kc_\delta}\right) = \frac{\delta_{\mathsf{g}}}{8\log n}$. It is easy to verify that $\text{UB1} \leq 2c\cdot\frac{\delta_g}{4}$, and $\text{UB2} \leq 2c\cdot\frac{\delta_g}{2}$.

We further set

$$\eta \triangleq \frac{\delta_{\mathsf{g}}}{16\cdot 32\cdot kt\cdot\log\frac{n}{\delta_{\mathsf{g}}}}, \tag{B.29}$$

and obtain $\text{UB3} = 2c\cdot\frac{\delta_g}{4}$.

Combining the upper bounds derived above for UB1, UB2, UB3, and by Equation (B.27), we get $f_{\text{AUX}}^{\text{avg}}(k') \leq 2c\cdot\delta_{\mathsf{g}}$. By Lemma 2 we know that the error probability is upper bounded by $p_{\text{ERR}} = \mathbb{E}_{\psi_{k'}}\left[p_{\text{ERR}}^{\text{MAP}}(\psi_{k'})\right] \leq \frac{f_{\text{AUX}}^{\text{avg}}(k')}{2c} \leq \delta_{\mathsf{g}}$. That is, with the cost k' specified in Equation (B.28), ECED is guaranteed to achieve $p_{\text{ERR}} \leq \delta_{\mathsf{g}}$.

It remains to compute the (exact) value of k'. Combining the definitions of $c \triangleq 8\left(\log(2n^2/\eta)\right)^2$ and $c_\delta \triangleq (6c+8)\log(n/\delta_{\mathsf{g}})$ with Equation (B.29) it is easy to verify that

$$c_\delta \leq c_1\cdot\left(\log\frac{nk}{\delta_{\mathsf{g}}}\right)^2\cdot\log\frac{n}{\delta_{\mathsf{g}}}$$

holds for some constant c_1. Therefore by Equation (B.28),

$$k' \leq k\cdot c_1\left(\log\frac{nk}{\delta_{\mathsf{g}}}\right)^2\log\frac{n}{\delta_{\mathsf{g}}}\cdot\frac{1}{c_\varepsilon}\ln\frac{8\log n}{\delta_{\mathsf{g}}} = O\left(\frac{k}{c_\varepsilon}\left(\log\frac{nk}{\delta_{\mathsf{g}}}\right)^2\left(\log\frac{n}{\delta_{\mathsf{g}}}\right)^2\right).$$

To put it in words, it suffices to run ECED for $O\left(\frac{k}{c_\varepsilon}\left(\log\frac{nk}{\delta_{\mathsf{g}}}\right)^2\left(\log\frac{n}{\delta_{\mathsf{g}}}\right)^2\right)$ steps to have expected error below δ_{g}, where k denotes the worst-case cost the optimal policy that achieves expected error probability $\delta_{\mathsf{opt}} \triangleq O\left(\frac{\delta_{\mathsf{g}}}{(\log n\cdot\log(1/\delta_{\mathsf{g}}))^2}\right)$; hence the completion of the proof. □

Appendix C: Examples when GBS and the most informative policy fail

In this section, we provide problem instances where GBS and/or the Most Informative Policy may fail, while ECED performs well. Since in the noise-free setting ECED is equivalent to EC^2, it suffices to demonstrate the limitations of GBS and the most informative policy, even if we provide just examples that apply to the noise-free setting.

C.1. A bad example for GBS*: Imbalanced equivalence classes*

We use the same example as provided in Golovin et al. [14]. Consider an instance with a uniform prior over n root-causes, $\theta_1, \ldots, \theta_n$, and two target values $y_1 = r(\theta_1) = \ldots r(\theta_{n-1})$, and $y_2 = r(\theta_n)$. There are tests $\mathcal{V} = \{1, \ldots, n\}$ such that $\mathbb{P}[X_e = 1 \mid \theta_i] = \mathbb{1}\{i = e\}$ (all of unit cost). Here, $\mathbb{1}\{\cdot\}$ is the indicator function. See Fig. 11 for illustration.

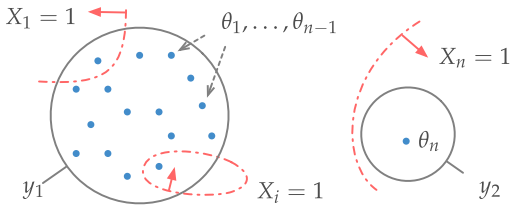


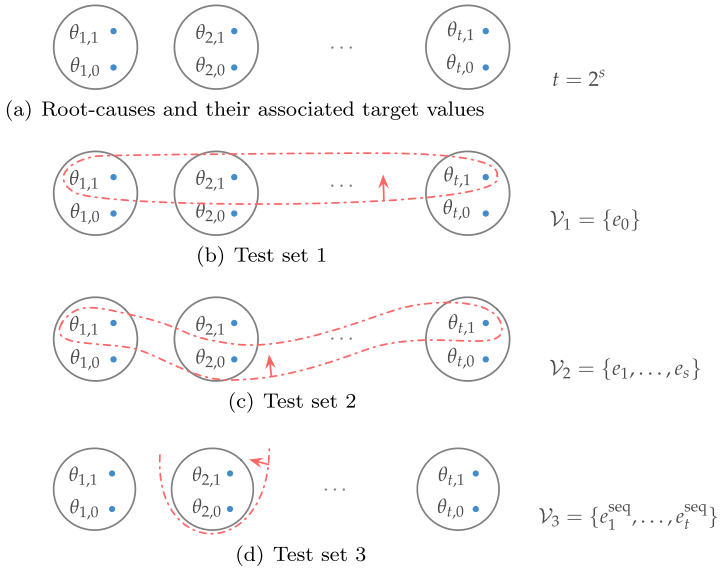
FIG 11. *A problem instance where GBS performs significantly worse than* ECED.

Now, suppose we want to solve Problem (2.1) for $\delta = 1/n$. Note that in the noise-free setting, the problem is equivalent to find a minimal cost policy π that achieves 0 prediction error, because once the error probability drops below $1/n$ we will know precisely which target value is realized.

In this case, the optimal policy only needs to select test n, however GBS may select tests $\{1, \ldots, n\}$ in order until running test e, where $\Theta = \theta_e$ is the true root-cause. Given our uniform prior, it takes $n/2$ tests in expectation until this happens, so that GBS pays, in expectation, $n/2$ times the optimal expected cost in this instance. Note that in this example, ECED (equivalently, EC^2) also selects test n, which is optimal.

C.2. A bad example for the most informative policy: Treasure hunt

In this section, we provide a *treasure-hunt* example, in which the most informative policy pays $\Omega(n/\log(n))$ times the optimal cost. This example is adapted from Golovin et al. [14], where they show that the most informative policy (referred to as the *Informative Gain* policy), as well as the myopic policy that greedily maximizes the reduction in the expected prediction error (referred as the *Value of Information* policy), both perform badly, compared with EC^2.



(a) Root-causes and their associated target values

(b) Test set 1

(c) Test set 2

(d) Test set 3

FIG 12. *A problem instance where the maximal informative policy, and the the myopic policy that greedily maximizes the reduction in the expected prediction error, perform significantly worse than* EC^2 *(equivalently,* ECED *in the noise-free setting).*

Consider the problem instance in Fig. 12(a). Fix $s > 0$ to be some integer, and let $t = |\mathcal{Y}| = 2^s$. For each target value $y_i \in \mathcal{Y}$, there exists two root-causes, i.e., $\theta_{i,1}$, $\theta_{i,0}$, such that $r(\theta_{i,1}) = r(\theta_{i,0}) = y_i$. Denote a root-cause as $\theta_{i,o}$, if it belongs to target i and is indexed by o. We assume a uniform prior over the root-causes: $\{\theta_{i,o}\}_{i\in\{1,\ldots,t\},o\in 0,1}$.

Suppose we want to solve Problem (2.1) for $\delta = 1/3$. Similarly with §C.1, the problem is equivalent to find a minimal cost policy π that achieves 0 prediction error, because once the error probability drops below 1/3, we will know precisely which target value is realized.

There are three set of tests, and all of them have binary outcomes and unit cost. The first set $\mathcal{V}_1 := \{e_0\}$ contains one test e_0, which tells us the value of o of the underlying root-cause $\theta_{i,o}$. Hence for all i, $\Theta = \theta_{i,o} \Rightarrow X_{e_0} = o$ (see Fig. 12(b)). The second set of tests are designed to help us quickly discover the index of the target value via binary search if we have already run e_0, but to offer no information whatsoever (in terms of expected reduction in the prediction error, or expected reduction in entropy of Y) if e_0 has not yet been run. There are a total number of s tests in the second set $\mathcal{V}_2 := \{e_1, e_2, \ldots, e_s\}$. For $z \in \{1, \ldots, t\}$, let $b_k(z)$ be the k^{th} least-significant bit of the binary encoding of z, so that $z = \sum_{k=1}^{s} 2^{k-1} b_k(z)$. Then, if $\Theta = \theta_{i,o}$, then the outcome of test $e_k \in \mathcal{V}_2$ is $X_{e_k} = \mathbb{1}\{\phi_k(i) = o\}$ (see Fig. 12(c)). The third set of tests are designed to allow us to do a (comparatively slow) sequential search on the index of the the target values. Specifically, we have $\mathcal{V}_3 := \{e_1^{\text{seq}}, \ldots, e_t^{\text{seq}}\}$, such that $\Theta = \theta_{i,o} \Rightarrow X_{e_k^{\text{seq}}} = \mathbb{1}\{i = k\}$ (Fig. 12(d)).

Now consider running the maximal informative policy π (the same analysis also applies to the value of information policy, which we omits from the paper). Note that in the beginning, no single test from $\mathcal{V}_1 \cup \mathcal{V}_2$ results in any change in the distribution over Y, as it remains uniform no matter which test is performed. Hence, the maximal informative policy only picks tests from $\mathcal{V}_3$, which have non-zero (positive) expected reduction in the posterior entropy of Y. In the likely event that the test chosen is not the index of Y, we are left with a residual problem in which tests in $\mathcal{V}_1 \cup \mathcal{V}_2$ still have no effect on the posterior. The only difference is that there is one less class, but the prior remains uniform. Hence our previous argument still applies, and π will repeatedly select tests in $\mathcal{V}_3$, until a test has an outcome of 1. In expectation, the cost of π is least $\text{cost}(\pi) \geq \frac{1}{t}\sum_{z=1}^{t} z = \frac{t+1}{2}$.

On the other hand, a smarter policy π^* will select test $e_0 \in \mathcal{V}_1$ first, and then performs a binary search by running tests $e_1, \ldots, e_s \in \mathcal{V}_2$ to determine $b_k(i)$ for all $1 \leq k \leq s$ (and hence to determine the index i of Y). Since the tests have unit cost, the cost of π^* is $\text{cost}(\pi^*) = s + 1$.

Since $t = 2^s$, and $n = 2t = 2^{s+1}$, we conclude that

$$\text{cost}(\pi) = \frac{t+1}{2} > \frac{t}{2} = \frac{n}{4}\frac{s+1}{\log n} = \frac{n}{4\log(n)}\text{cost}(\pi^*).$$

Acknowledgments

This work was supported in part by ERC StG 307036, a Microsoft Research Faculty Fellowship, and a Google European Doctoral Fellowship.

References

[1] Maria-Florina Balcan and Ruth Urner. Active learning–modern learning theory. *Encyclopedia of Algorithms*, 2015.

[2] G. Bellala, S. Bhavnani, and C. Scott. Extensions of generalized binary search to group identification and exponential costs. In *NIPS*, 2010.

[3] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. MR0070925

[4] V. T. Chakaravarthy, V. Pandit, S. Roy, P. Awasthi, and M. Mohania. Decision trees for entity identification: Approximation algorithms and hardness results. In *SIGMOD/PODS*, 2007.

[5] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. MR1390519

[6] Yuxin Chen and Andreas Krause. Near-optimal batch mode active learning and adaptive submodular optimization. In *ICML*, 2013.

[7] Yuxin Chen, S. Hamed Hassani, Amin Karbasi, and Andreas Krause. Sequential information maximization: When is greedy near-optimal? In *COLT*, 2015.

[8] Yuxin Chen, Shervin Javdani, Amin Karbasi, James Andrew Bagnell, Siddhartha Srinivasa, and Andreas Krause. Submodular surrogates for value of information. In *AAAI*, 2015.
[9] S. Dasgupta. Analysis of a greedy active learning strategy. In *NIPS*, 2004.
[10] Sanjoy Dasgupta. Analysis of a greedy active learning strategy. In *NIPS*, 2004.
[11] Amol Deshpande, Lisa Hellerstein, and Devorah Kletenik. Approximation algorithms for stochastic boolean function evaluation and stochastic submodular set cover. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1453–1467. SIAM, 2014. MR3376467
[12] Valerii Vadimovich Fedorov. *Theory of optimal experiments*. Elsevier, 1972. MR0403103
[13] Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *JAIR*, 2011. MR2874807
[14] Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. In *NIPS*, 2010.
[15] Steve Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007. MR2930645
[16] Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2–3):131–309, 2014.
[17] David Heckerman, John Breese, and Koos Rommelse. Troubleshooting under uncertainty. Technical report, Technical Report MSR-TR-94-07, Microsoft Research, 1994.
[18] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl. An algorithmic framework for performing collaborative filtering. In *SIGIR*, 1999.
[19] R.A. Howard. Information value theory. *Systems Science and Cybernetics, IEEE Trans. on*, 2(1):22–26, 1966.
[20] Matti Kääriäinen. Active learning in the non-realizable case. In *Algorithmic Learning Theory*, pages 63–77, 2006. MR2324114
[21] Haim Kaplan, Eyal Kushilevitz, and Yishay Mansour. Learning with attribute costs. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 356–365. ACM, 2005.
[22] Richard M Karp and Robert Kleinberg. Noisy binary search and its applications. In *SODA*, 2007.
[23] S Rao Kosaraju, Teresa M Przytycka, and Ryan Borgstrom. On an optimal split tree problem. In *Algorithms and Data Structures*, pages 157–168. Springer, 1999.
[24] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *NIPS*, 2012.
[25] Robert Nowak. Noisy generalized binary search. In *NIPS*, 2009.
[26] Debajyoti Ray, Daniel Golovin, Andreas Krause, and Colin Camerer. Bayesian rapid optimal adaptive design (broad): Method and application distinguishing models of risky choice. *Tech. Report*, 2012.

[27] M. C. Runge, S. J. Converse, and J. E. Lyons. Which uncertainty? using expert elicitation and expected value of information to design an adaptive program. *Biological Conservation*, 2011.

[28] B. Settles. *Active Learning*. Morgan & Claypool, 2012.

[29] Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *AISTATS*, 2015.

[30] William F. Sharpe. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance*, 1964.

[31] Richard D Smallwood and Edward J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

[32] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 1992.

[33] P.P. Wakker. *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press, 2010. MR2829560

[34] Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems*, pages 442–450, 2014.