JMLR: Workshop and Conference Proceedings vol 40:1–26, 2015

Sequential Information Maximization: When is Greedy Near-optimal?

Yuxin Chen† YUXIN.CHEN@INF.ETHZ.CH
S. Hamed Hassani† HAMED@INF.ETHZ.CH
Amin Karbasi‡ AMIN.KARBASI@YALE.EDU
Andreas Krause† KRAUSEA@ETHZ.CH
† *ETH Zürich, Zürich, Switzerland*
‡ *Yale University, New Haven, CT, USA*

Abstract

Optimal information gathering is a central challenge in machine learning and science in general. A common objective that quantifies the usefulness of observations is Shannon's mutual information, defined w.r.t. a probabilistic model. Greedily selecting observations that maximize the mutual information is the method of choice in numerous applications, ranging from Bayesian experimental design to automated diagnosis, to active learning in Bayesian models. Despite its importance and widespread use in applications, little is known about the theoretical properties of sequential information maximization, in particular under noisy observations. In this paper, we analyze the widely used greedy policy for this task, and identify problem instances where it provides provably near-maximal utility, even in the challenging setting of persistent noise. Our results depend on a natural separability condition associated with a channel injecting noise into the observations. We also identify examples where this separability parameter is necessary in the bound: if it is too small, then the greedy policy fails to select informative tests.

Keywords: Active learning, Information theory, Optimization

1. Introduction

Optimal information gathering, i.e., selectively acquiring most useful data, is a task of central importance in machine learning and science in general. Many such problems can be formalized as sequentially selecting *tests* – variables to observe in a probabilistic model – in order to maximally reduce the uncertainty about a target variable (often called *hypothesis*) of interest Y. This setting includes Bayesian experimental design, as originally studied by Lindley (1956), where tests correspond to experiments that can be carried out, and the target variable encodes model parameters of interest. It naturally maps to applications such as medical diagnosis (performing medical tests that are most informative about the patient's condition (Berry et al., 2010)), sensor selection (selecting informative sensors to query, e.g., for target detection (Williams et al., 2007)), and numerous others. In machine learning, it is related to the task of active learning in Bayesian models (where tests correspond to obtaining labels for data points, and the target variable corresponds to unknown model parameters). In these applications, tests are usually expensive, and we seek maximal information subject to a constraint on the cost. A widely used notion of *informativeness* is given by the reduction of Shannon entropy (Shannon, 1948), also known as the *mutual information*, about the target variable.

Maximizing the mutual information between sets of random variables has a rich history in machine learning (Luttrell, 1985; MacKay, 1992). It is perhaps best understood in the *a priori selection* (a.k.a. open loop, or non-adaptive) setting, where the set of all tests to be executed is determined

ahead of time, i.e., before any observations have been made. It is known that the problem of selecting a set of most informative variables of restricted cardinality is generally NP-hard (Ko et al., 1995). By leveraging the theory of *submodular functions* (Nemhauser et al., 1978), Krause and Guestrin (2005) showed that under some conditions, near-optimal solutions can be identified efficiently. In particular, under the assumption that the test outcomes are conditionally independent given the target variable Y, the mutual information between selected test sets and Y is submodular, and therefore a simple greedy algorithm leads to a $1 - 1/e$ approximation of the optimal solution.

However, in many applications, it is more natural to consider *sequential* (a.k.a. closed-loop, or adaptive) selection. Here, one considers *policies* (decision trees, conditional plans) that select the next test to carry out depending on observations made by previous tests. This sequentiality provides an informational advantage, allowing in general to obtain more information than committing to all tests ahead of time. A natural policy that finds widespread use in practice is the *most informative selection policy* that in each step greedily picks the test that provides the maximal reduction in uncertainty (quantified in terms of Shannon entropy) about the target variable. Despite its widespread use, not much is known about the theoretical properties of this greedy policy, in particular in the practically important setting where observations are noisy. A general framework to study the performance of greedy policies is *adaptive submodularity* (Golovin and Krause, 2011). It is known that if a sequential problem is adaptive submodular, then optimizing it greedily results in near-optimal performance. Unfortunately, the mutual information criterion violates the adaptive submodularity condition, and hence does not fall into this framework. In the case where tests are *noise-free* (i.e., their outcome is a deterministic function of Y), it is known that greedily optimizing mutual information is effective (Dasgupta, 2005; Zheng et al., 2005). If tests are noisy, but can be repeated with i.i.d. outcomes, the problem can effectively be reduced to the noise-free setting at small increase in cost[1] (Kääriäinen, 2006; Naghshvar et al., 2013a).

However, in many applications, noise is *persistent*: Repeating a test is impossible, or will produce identical observations. Experiments, for example, might be systematically biased due to environmental conditions. Or experts, labeling examples in active learning make consistent mistakes. For such practically highly relevant settings, to the best of our knowledge nothing is known about the performance of the most informative selection policy.

Our contribution In this paper, we establish the first rigorous information-theoretic analysis of the most informative selection policy that holds even under persistent noise. Specifically, we consider a general probabilistic model, where the originally deterministic tests are corrupted by some arbitrary noisy channel. We derive a lower bound on the utility achieved by the greedy policy in terms of a *channel separability condition* (see, Definition 1), a simple measure that characterizes the severity of noise. We further provide an example to show that such measure is important in the bound. It follows from our results that under common assumptions made about the noise (e.g., binary symmetric channel), the sequential information maximization criterion behaves near-optimally. Hence our results theoretically justify why the mutual information criterion has been found to be effective in these settings. Our analysis also sheds light on cases where greedy information maximization may fail, and thus nonmyopic policies, e.g., using look-ahead, might be required.

1. Simply repeat the test $O(\log \frac{1}{\delta})$ times, until the most likely outcome is determined with probability $1 - \delta$.

2. Related Work

Optimal Information Gathering. The general problem of optimal information gathering has been studied in diverse fields, including active learning (MacKay, 1992; Dasgupta and Langford, 2011; Settles, 2012), experimental design (Lindley, 1956; Fedorov, 1972), evaluation of (stochastic) Boolean functions (Kaplan et al., 2005; Deshpande et al., 2014), channel coding with feedback (Horstein, 1963; Burnashev, 1976), and active hypothesis testing (Chernoff, 1959; Nowak, 2009; Naghshvar et al., 2013b). It has been theoretically studied in two settings 1) the information maximization problem, where the goal is to maximize utility under a budget constraint, and 2) the cost minimization problem, where the aim is to achieve a target utility with minimal cost. In this paper we focus our analysis on the first variant, and derive a lower bound on the utility achievable by the most informative selection policy.

Active Learning in the Noise-free Setting. In machine learning, information gathering has been mainly studied in *active learning*, where the goal is to (sequentially) query labels for data points that most effectively reduce the uncertainty about an underlying hypothesis. Settings studied include the *stream-based setting*, where unlabeled data points arrive one at a time; and the *pool-based setting*, where a large collection of unlabeled data is available for querying.

The sequential information maximization problem we study is most closely related to pool-based active learning with a Bayesian prior on a (finite) set of hypotheses. Assuming binary labels that are noise free (i.e., the realizable setting), each observation eliminates all inconsistent hypotheses, and one seeks to determine the correct hypothesis while minimizing the cost of testing. Finding the optimal policy is NP-hard in this setting (Chakaravarthy et al., 2007), but a simple greedy algorithm, *generalized binary search* (GBS), is guaranteed to be competitive with the optimal policy (Freund et al., 1997; Kosaraju et al., 1999; Dasgupta, 2005; Golovin and Krause, 2011): at each step, GBS chooses the test that splits the hypothesis space as evenly as possible, and the expected number of tests is within a factor of $O(\log n)$ of the optimal policy's cost. In fact, one can show that under the noise-free setting, GBS is equivalent to the most informative selection policy.

Active Learning with Noisy Observations. Moving beyond the noise-free setting, active learning has been analyzed under various noise models under the framework of statistical learning theory. In the stream-based setting, policies are known that work even in the agnostic setting (Balcan et al., 2006; Hanneke, 2007). The stream-based setting forfeits control over the sequence of examples considered, which empirically can lead to worse performance compared to more aggressive pool-based methods (Gonen et al., 2013). On the other hand, current theoretical results for the pool-based setting are restricted to limited hypotheses classes (e.g., halfspaces as in Balcan et al. (2007); Gonen et al. (2013)) and restricted noise assumptions (e.g., Tsybakov (2004); Hanneke and Yang (2014)).

Note that our result is orthogonal to most existing theoretical results in active learning that establish label complexity bounds: in active learning one usually aims to minimize the cost, while guaranteeing prediction accuracy; whereas in this paper, we seek *computationally-efficient* approaches that are *provably competitive* with the optimal policy in terms of maximizing the utility. In most active learning literature (e.g., Dasgupta (2005); Hanneke (2007), Hanneke (2014); Balcan and Urner (2015)), the results have been characterized in terms of the structure of the hypotheses class, as well as additional distribution-dependent complexity measures. In contrast, we do not need to bound how the optimal policy behaves, and hence we make no assumptions on the hypothesis class. Rather, our (near-optimality) bound only depends on properties of the channel injecting the noise.

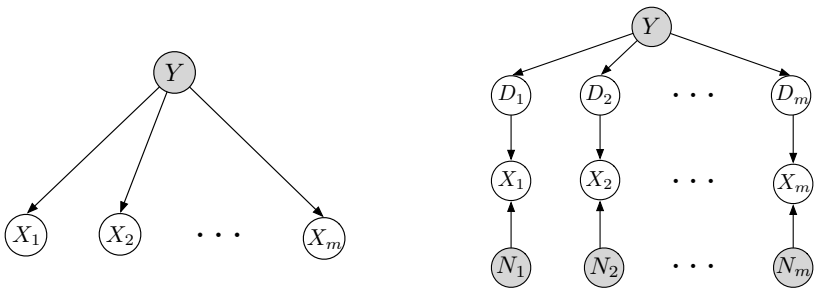
Perhaps most similar to our approach is the recent work of Golovin et al. (2010) and Chen et al. (2015), who have used the adaptive submodularity framework to obtain efficient greedy algorithms with provable (logarithmic) approximation guarantees for active learning with persistent noise. There, however, the problem is restricted to a bounded noise setting, i.e., it only allows a bounded number of flips on the test outcomes, and the results degrade with the support of the noisy channel (Golovin et al., 2010). We relax the bounded noise assumption, and show that a near-optimal bound on the greedy algorithm holds for more general information maximization problems.

3. Preliminaries

We now introduce notation, our basic model and the formal problem statement.

3.1. Basic Model

The basic model that we consider is as follows: We are given a hidden random variable Y that ranges among a set $\mathcal{Y} = \{y_1, \cdots, y_n\}$ with some known distribution $Y \sim \Pr(Y = y)$. The goal is to learn the value of Y from (a subset of) observable discrete random variables $X_1, \cdots, X_m$ statistically dependent on Y. From now on, we think of the value of Y as representing a true "hypothesis" among a set of possible n hypotheses and each of the X_e's as a "test" that we could perform, whose observation reveals some information about the true hypothesis. Here e is the indexing variable of the test, i.e., $e \in [m]$. Denote the observed outcome of X_e as x_e, and assume that each test has unit cost. We further adopt the common assumption in Bayesian experimental design that the joint probability distribution $P(Y, X_1, \ldots, X_m)$ is known, and that we can perform efficient inference (i.e., can compute marginal and conditional distributions). Our goal is to sequentially (adaptively) choose a set of k' tests that are maximally informative about Y. An important assumption we make is that X_e's are conditionally independent given Y (see Figure 1(*a*)). Equivalently, we assume that the each test X_e depends on the hidden variable Y and another independent latent variable, called the *noise* N_e, in the following way: First, Y goes through a deterministic mapping $D_e := D_e(Y)$, i.e, each D_e is a function of Y. The output of D_e will then be perturbed by the noise N_e, and produce the test (observable random variable) X_e (see Figure 1(*b*)). Hence X_e is a deterministic function of the noise N_e and D_e.



(*a*) (*b*)

Figure 1: (a) The Naïve Bayes model. (b) An equivalent Bayes net representation

Example 1 *An example for our setting is the generalized binary search (GBS) problem, where Y represents a randomly chosen hypothesis and each X_e is a binary random variable, representing the*

binary label of example e under hypothesis Y. In the noise-free setting, $X_e = D_e$ is a deterministic function of Y. In the noisy setting, X_e results from flipping the (deterministic) outcome of D_e with probability ϵ and the flipping events of the tests are independent. In other words, we can write $X_e = D_e \oplus N_e$, where D_e is the true label and a deterministic function of Y, N_e is a binary random variable with $\Pr(N_e = 1) = \epsilon$, and $\oplus$ denotes the addition in $\mathbb{F}_2 = \{0, 1\}$ (i.e., the XOR operation).

3.2. Policies

We consider adaptive strategies for picking the tests. From now on, we encode an adaptive strategy as a policy π. In words, a policy π specifies which test to pick[2] next based on the the tests picked so far and their corresponding outcomes. We consider policies of fixed length, say k. Hence, upon completion, policy π returns a sequence of k test-outcome pairs denoted by ψ_π, i.e., $\psi_\pi \triangleq \{(e_{\pi,1}, x_{e_{\pi,1}}), (e_{\pi,2}, x_{e_{\pi,2}}), \cdots, (e_{\pi,k}, x_{e_{\pi,k}})\}$. Note that what π returns in the end is itself random, dependent on the (random) outcomes of the selected tests (as well as the decisions that π has made). Once ψ_π is observed, we obtain a new posterior of Y, and hence the associated entropy $\mathbb{H}(Y \mid \psi_\pi)$. We define the entropy of Y given the policy π as follows

$$\mathbb{H}(Y \mid \pi) \triangleq \mathbb{E}_{\psi_\pi}[\mathbb{H}(Y \mid \psi_\pi)]. \tag{1}$$

In words, $\mathbb{H}(Y \mid \pi)$ is the expected entropy of the posterior of Y given the final outcome of π. Also, the mutual information between π and Y is

$$\mathbb{I}(\pi; Y) = \mathbb{H}(Y) - \mathbb{H}(Y \mid \pi), \tag{2}$$

which indicates the expected amount of information that π provides about Y upon completion.

We then define the optimal policy $\pi_{\mathsf{OPT}[k]}$ to be the policy that achieves the maximal expected mutual information, i.e.,

$$\pi_{\mathsf{OPT}[k]} = \operatorname*{argmax}_{\pi \in \Pi_{[k]}} \mathbb{I}(\pi; Y), \tag{3}$$

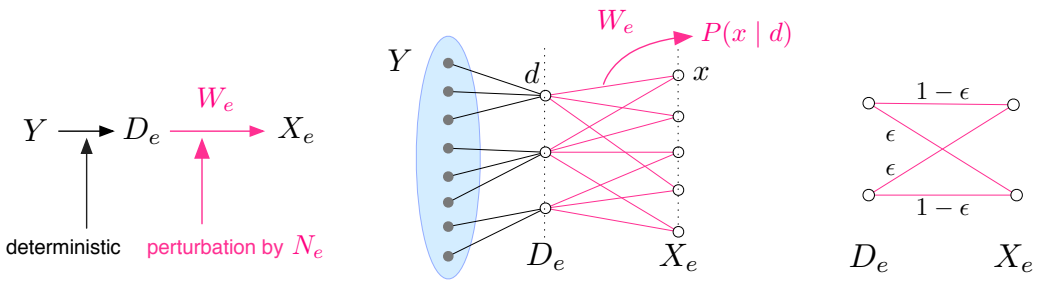
where $\Pi_{[k]}$ is the set of all policies of length k. Note that computing the optimal policy is intractable in general. A very well known, efficient and intuitive policy is the one that greedily picks the test that reduces the current entropy of Y the most, or equivalently, has the maximum mutual information w.r.t. the current distribution of Y. Denote this *most informative selection policy* of length k by $\pi_{\mathsf{Greedy}[k]}$. Formally speaking, $\pi_{\mathsf{Greedy}[k]}$ operates as follows: At any round $\ell + 1$, $0 \leq \ell \leq k - 1$, a test $X_{e_{\ell+1}}$, will be picked according to what has previously been observed, i.e., $\psi_\ell \triangleq \{(e_1, x_{e_1}), \ldots, (e_{\ell-1}, x_{e_{\ell-1}})\}$. We have

$$e_{\ell+1} = \operatorname*{argmax}_{e \in [m]} \mathbb{I}(X_e; Y \mid \psi_\ell). \tag{4}$$

3.3. Channel Induced by Noise

As explained above, for any $e \in [m]$ the random variable D_e is a deterministic function of Y. The value of D_e is then perturbed by the noise (N_e) to generate the test variable X_e. Note that the perturbation of N_e is assumed to take place independently of Y and hence can be characterized through a conditional probability distribution $\Pr(X_e = x \mid D_e = d)$ where $x \in \mathcal{X}_e$, $d \in \mathcal{D}_e$, and

2. We also allow randomized policies where the next test can be picked according to some probability distribution which possibly depends on what has been observed so far.



(*a*) The Channel (*b*) Illustration of the noise channel (*c*) BSC

Figure 2: Illustration of the channel induced by noise. (a-b) show the data generation process. In (c) we illustrate the binary symmetric channel for Example 1.

$\mathcal{D}_e, \mathcal{X}_e$ are the support of D_e, X_e. We refer to this conditional probability distribution as *the channel induced by the noise* and denote it by W_e (see Figure 2).

The test X_e depends on Y only through D_e, i.e., we have the Markov chain $Y \to D_e \to X_e$. As a result,

$$\mathbb{I}(X_e; Y) = \mathbb{I}(X_e; D_e),$$

and the latter is by definition less than the capacity of the channel W_e. We now introduce another parameter for the channel W_e and will later establish its importance for sequential selection.

Definition 1 (Separability of a channel) *Consider a channel W with associated conditional probability distribution $\{p(x \mid d)\}_{d\in\mathcal{D}, x\in\mathcal{X}}$. Note that given each $d \in \mathcal{D}$, $p(\cdot \mid d)$ is a probability distribution over $\mathcal{X}$. The separability of W, denoted by $S(W)$, is then defined by*

$$S(W) = \left(\min_{d,d'\in\mathcal{D}: d\neq d'} \left| p(\cdot \mid d) - p(\cdot \mid d') \right|_{\mathrm{TV}} \right)^2. \tag{5}$$

Here, $|\cdot|_{\mathrm{TV}}$ denotes the total variation distance. Also, if $|\mathcal{D}| = 1$ we let $S(W) = 1$. Intuitively, for a channel W and two inputs d, d', the value $|p(\cdot \mid d) - p(\cdot \mid d')|_{\mathrm{TV}}$ is an indicator of how much the channel can differentiate between d and d'. E.g., if $|p(\cdot \mid d) - p(\cdot \mid d')|_{\mathrm{TV}} = 0$ then $p(\cdot \mid d) = p(\cdot \mid d')$, in which case it is impossible to distinguish d from d' given the output of the channel. On the other hand, if $|p(\cdot \mid d) - p(\cdot \mid d')|_{\mathrm{TV}} = 1$ then from the output we can for sure exclude either d or d' (i.e., if we know that the input was either d or d', then we can say from the output which one is the input).

As mentioned above, for any $e \in [m]$ we have an associated channel W_e which is induced by the noise N. We denote by $S_{\min}$ the minimum value of separability over all the channels W_e, i.e.,

$$S_{\min} = \min_{e\in[m]} S(W_e).$$

In the noisy GBS example (Example 1), it is easy to see that the separability of the binary symmetric channel (see Figure 2-(c)) is $S_{\min} = (1 - 2\epsilon)^2$.

4. Main Result

We are now ready to state our main result, which provides the first approximation bound on the performance of the most informative selection policy under persistent noise.

Theorem 2 *Consider the sequential information maximization problem, where we run the most informative selection policy $\pi_{\textsf{Greedy}}$ till length k'. For any $\delta > 0$ and $k \in \mathbb{N}$, we have*[3]

$$\mathbb{I}\left(\pi_{\textit{Greedy}[k']}; Y\right) \geq \left(\mathbb{I}\left(\pi_{\textit{OPT}[k]}; Y\right) - \delta\right)\left(1 - \exp(-\frac{k'}{k\gamma \max\{\log n, \log \frac{1}{\delta}\}})\right), \tag{6}$$

where $n = |\mathcal{Y}|$ is the number of possible values of Y, and γ is a constant that only depends on the noise N, concretely: $\gamma = \frac{7}{S_{\min}}$.

We present the proof of the Theorem in Section 5. Here, we list a few noteworthy observations regarding Theorem 2. First, suppose that for some fixed $0 < \alpha < 1$ we have that $\delta = \alpha \mathbb{I}\left(\pi_{\textsf{OPT}[k]}; Y\right)$. Thus, δ is expressed as a fraction of the maximum mutual information obtainable by any policy. Then the RHS of Inequality (6) turns into a multiplicative bound in terms of α:

$$\mathbb{I}\left(\pi_{\textsf{Greedy}[k']}; Y\right) \geq \mathbb{I}\left(\pi_{\textsf{OPT}[k]}; Y\right)(1-\alpha)\left(1 - \exp\left(-\frac{k'}{k\gamma \max\{\log n, \log \frac{1}{\alpha \mathbb{I}\left(\pi_{\textsf{OPT}[k]}; Y\right))}\}}\right).$$

We note that for many reasonable scenarios, $\mathbb{I}\left(\pi_{\textsf{OPT}[k]}; Y\right)$ is typically at least a few bits, otherwise arguably the information gathering task is ill-posed / infeasible. In this case, if we assume $\mathbb{I}\left(\pi_{\textsf{OPT}[k]}; Y\right) \geq 1$, then we obtain a lower bound where the multiplicative factor only depends on the noise channel:

$$\mathbb{I}\left(\pi_{\textsf{Greedy}[k']}; Y\right) \geq \mathbb{I}\left(\pi_{\textsf{OPT}[k]}; Y\right)(1-\alpha)\left(1 - \exp(-\frac{k'}{k\gamma \max\{\log n, \log \frac{1}{\alpha}\}})\right).$$

Another way to interpret the result is to use the fact that $I(\pi_{\textsf{OPT}[k]}; Y) \leq \log n$. From (6) we obtain

$$\mathbb{I}\left(\pi_{\textsf{Greedy}[k']}; Y\right) \geq \mathbb{I}\left(\pi_{\textsf{OPT}[k]}; Y\right) - \delta - \log n\left(1 - \exp(-\frac{k'}{k\gamma \max\{\log n, \log \frac{1}{\delta}\}})\right),$$

As a consequence, if we choose $k' \geq k\gamma \max\{\log n, \log \frac{1}{\delta}\} \ln(\frac{\log n}{\delta})$, then we have

$$\mathbb{I}\left(\pi_{\textsf{Greedy}[k']}; Y\right) \geq \mathbb{I}\left(\pi_{\textsf{OPT}[k]}; Y\right) - 2\delta.$$

Hence, we can get arbitrarily close – up to δ in absolute terms – to the optimal mutual information achievable within k tests by greedily selecting k' tests, which is within a logarithmic factor (in terms of $\log n$ and $\log \frac{1}{\delta}$) of k.

Discussion. A few comments are in order. First, as an example, for the GBS problem in Example 1, $\gamma = \frac{7}{(1-2\epsilon)^2}$, and the lower bound we get for the greedy algorithm is $\mathbb{I}\left(\pi_{\textsf{Greedy}[k']}; Y\right) \geq \left(\mathbb{I}\left(\pi_{\textsf{OPT}[k]}; Y\right) - \delta\right)\left(1 - \exp(\frac{k'(1-2\epsilon)^2}{7k \max\{\log n, \log \frac{1}{\delta}\}})\right)$.

Second, in Appendix C, we construct an example, where the ratio between the gain of $\pi_{\textsf{Greedy}[k]}$ and the optimal policy $\pi_{\textsf{OPT}[k]}$ is at most $c_0 S_{\min}$, where c_0 is some constant. However, for our example to hold, we actually require that $S_{\min}$ to be at least $\Omega\left(1/\log n\right)$. On the other hand, both $S_{\min}$ and $1/\log n$ are involved in the lower bound in Theorem 2. It remains an open problem to decide which combination of $S_{\min}$ and $1/\log n$ is indeed necessary in the lower bound.

Third, the $S_{\min}$ involved in our bound is defined over all the possible tests *picked by* $\pi_{\textsf{Greedy}}$ or $\pi_{\textsf{OPT}}$. Therefore, if there are some tests which are "purely noisy", i.e., the separability of their associated noise channels have $S(W) = 0$, then clearly both $\pi_{\textsf{Greedy}}$ and $\pi_{\textsf{OPT}}$ will disregard those tests, and hence their $S(W)$'s don't affect our lower bound.

3. In this paper all the log's are in base 2

5. Proofs

In this section we prove Theorem 2. A key lemma in proving the theorem is as follows.

Lemma 3 *Consider the probabilistic model of Figure 1 with an arbitrary probability distribution $Pr(\cdot)$. Also, consider any adaptive policy π which chooses k tests among $\{X_e\}_{e\in[m]}$ and gains mutual information $\mathbb{I}(\pi;Y)$. Then, we must have*

$$\max_{e\in[m]} \mathbb{I}(X_e;Y) \geq \frac{\mathbb{I}(\pi;Y)}{k\gamma \max\left\{\log n, \log \frac{1}{\mathbb{I}(\pi;Y)}\right\}}. \tag{7}$$

We relegate the proof of this lemma to the the next section. Let us now see how the result of Theorem 2 follows from this lemma.

Proof [Proof of Theorem 2] Now, we show that Eq. (6) holds for any for any policy $\pi_{[k]}$ of length k. Let Ψ_ℓ be a random variable representing the first ℓ tests (and their associated outcomes) that have been selected by the greedy policy π_{Greedy}, and ψ_ℓ be a specific realization of Ψ_ℓ. In the decision tree representation of π_{Greedy}, ψ_ℓ represents a path from the root to a node at level ℓ (see Figure 3). Now suppose we have run the greedy policy π_{Greedy} till level ℓ, and have observed the realized path ψ_ℓ (thus ψ_ℓ is a sequence of ℓ chosen tests and their observed outcomes). At this point, the greedy algorithm picks a new test according to the rule (4). Therefore, the expected gain of the greedy algorithm at time $\ell+1$ is $\max_{e\in[m]} \mathbb{I}(X_e;Y \mid \psi_\ell)$.

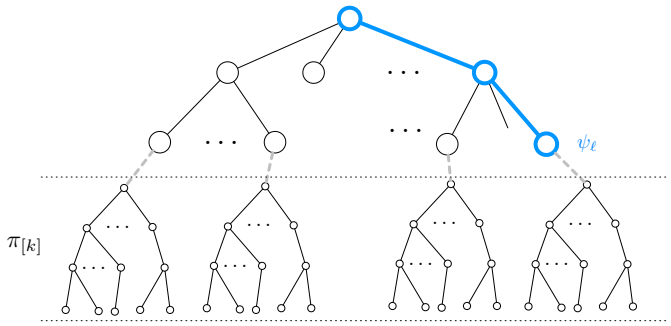


Figure 3: The decision tree representation of policies (i) π_{Greedy} of length ℓ, and (ii) $\pi_{[k]}$ of length k. After π_{Greedy} has selected ℓ tests (observed ψ_ℓ), we run policy $\pi_{[k]}$, as if from fresh start.

Let us now consider the following thought experiment. Assume the same setting as above (i.e., we have observed ψ_ℓ) and we run the policy $\pi_{[k]}$ of length k as if from a fresh start[4] (Figure 3), i.e., $\pi_{[k]}$ is run by totally neglecting the observation ψ_ℓ. The policy $\pi_{[k]}$ then outputs a realization ψ_π. The expected information we obtain by using the aforementioned version of π (that totally neglects the observation ψ_ℓ) is $\mathbb{H}(Y \mid \psi_\ell) - \mathbb{H}\left(Y \mid \psi_\ell, \pi_{[k]}\right)$ or equivalently $\mathbb{I}\left(\pi_{[k]}; Y \mid \psi_\ell\right)$. We can now use the result of Lemma 3 to relate the gain of the greedy to the gain of $\pi_{[k]}$. An important point to note here is that the result of Lemma (3) holds for *any* probability distribution on the Bayesian network of Figure 1. In particular, by conditioning all our distributions on the observation ψ_ℓ, and

4. This is indeed known as the concatenation of the two policies π_{Greedy} and π, see Golovin and Krause (2011).

by using Lemma 3, we obtain

$$\max_{e\in[m]} \mathbb{I}\left(X_e; Y \mid \psi_\ell\right) \geq \mathbb{E}_{\psi_\ell}\left[\frac{\mathbb{I}\left(\pi_{[k]}; Y \mid \psi_\ell\right)}{k\gamma \max\{\log n, \log \frac{1}{\mathbb{I}\left(\pi_{[k]}; Y \mid \psi_\ell\right)}\}}\right]. \tag{8}$$

Now, by further averaging over ψ_ℓ, the expected entropy reduction by running $\pi_{[k]}$ after $\pi_{\mathsf{Greedy}[\ell]}$ is

$$\begin{aligned}
\mathbb{E}_{\psi_\ell}\left[\mathbb{I}\left(\pi_{[k]}; Y \mid \psi_\ell\right)\right] &= \mathbb{E}_{\psi_\ell}\left[\mathbb{H}\left(Y \mid \psi_\ell\right) - \mathbb{H}\left(Y \mid \pi_{[k]}, \psi_\ell\right)\right] \\
&= \mathbb{H}\left(Y \mid \pi_{\mathsf{Greedy}[\ell]}\right) - \mathbb{H}\left(Y \mid \pi_{[k]}, \pi_{\mathsf{Greedy}[\ell]}\right) \\
&\geq \mathbb{H}\left(Y \mid \pi_{\mathsf{Greedy}[\ell]}\right) - \mathbb{H}\left(Y \mid \pi_{[k]}\right) \\
&= \mathbb{I}\left(\pi_{[k]}; Y\right) - \mathbb{I}\left(\pi_{\mathsf{Greedy}[\ell]}; Y\right)
\end{aligned} \tag{9}$$

Note here that $\mathbb{I}\left(\pi_{[k]}; Y\right)$ is the total information gain of the policy $\pi_{[k]}$ about Y. Fix $\delta > 0$, and denote $\alpha := k\gamma \max\{\log n, \log \frac{1}{\delta}\}$. We can resume (8) as follows

$$\begin{aligned}
\mathbb{E}_{\psi_\ell}\left[\max_{e\in E} \mathbb{I}\left(X_e; Y \mid \psi_\ell\right)\right] &\geq \mathbb{E}_{\psi_\ell}\left[\frac{\mathbb{I}\left(\pi_{[k]}; Y \mid \psi_\ell\right)}{k\gamma \max\{\log n, \log \frac{1}{\mathbb{I}\left(\pi_{[k]}; Y \mid \psi_\ell\right)}\}}\right] \\
&\geq \mathbb{E}_{\psi_\ell}\left[\frac{\mathbb{I}\left(\pi_{[k]}; Y \mid \psi\right) \cdot \mathbb{1}\left\{\mathbb{I}\left(\pi_{[k]}; Y \mid \psi_\ell\right) > \delta\right\}}{k\gamma \max\{\log n, \log \frac{1}{\mathbb{I}\left(\pi_{[k]}; Y \mid \psi_\ell\right)}\}}\right] \\
&\geq \frac{1}{\alpha}\left(\mathbb{E}_{\psi_\ell}\left[\mathbb{I}\left(\pi_{[k]}; Y \mid \psi_\ell\right) \cdot \mathbb{1}\left\{\mathbb{I}\left(\pi_{[k]}; Y \mid \psi_\ell\right) > \delta\right\} + \delta\right] - \delta\right) \\
&\geq \frac{1}{\alpha}\left(\mathbb{E}_{\psi_\ell}\left[\mathbb{I}\left(\pi_{[k]}; Y \mid \psi_\ell\right)\right] - \delta\right) \\
&\overset{\text{(inequality (9))}}{\geq} \frac{1}{\alpha}\left(\mathbb{I}\left(\pi_{[k]}; Y\right) - \mathbb{I}\left(\pi_{\mathsf{Greedy}[\ell]}; Y\right) - \delta\right)
\end{aligned}$$

Rearranging the terms, we have

$$\begin{aligned}
\mathbb{I}\left(\pi_{[k]}; Y\right) - \delta - \mathbb{I}\left(\pi_{\mathsf{Greedy}[\ell]}; Y\right) &\leq \alpha \mathbb{E}_{\psi_\ell}\left[\max_{e\in E} \mathbb{I}\left(X_e; Y \mid \psi_\ell\right)\right] \\
&= \alpha\left(\mathbb{H}\left(Y \mid \pi_{\mathsf{Greedy}[\ell]}\right) - \mathbb{E}_{\psi_\ell}\left[\min_{e\in E} \mathbb{H}\left(Y \mid X_e, \psi_\ell\right)\right]\right) \\
&= \alpha\left(\mathbb{I}\left(\pi_{\mathsf{Greedy}[\ell+1]}; Y\right) - \mathbb{I}\left(\pi_{\mathsf{Greedy}[\ell]}; Y\right)\right)
\end{aligned} \tag{10}$$

Let $\Delta_\ell := \mathbb{I}\left(\pi_{[k]}; Y\right) - \delta - \mathbb{I}\left(\pi_{\mathsf{Greedy}[\ell]}; Y\right)$, so that Inequality (10) implies $\Delta_\ell \leq \alpha \times (\Delta_\ell - \Delta_{\ell+1})$. From here we get $\Delta_{\ell+1} \leq \left(1 - \frac{1}{\alpha}\right)\Delta_\ell$, and hence $\Delta_{[k']} \leq \left(1 - \frac{1}{\alpha}\right)^{k'} \Delta_0 \leq \exp\left(-\frac{k'}{\alpha}\right)\Delta_0$. Thus, $\mathbb{I}\left(\pi_{[k]}; Y\right) - \delta - \mathbb{I}\left(\pi_{\mathsf{Greedy}[k']}; Y\right) < \exp\left(-\frac{k'}{\alpha}\right)\Delta_0 \leq \exp\left(-\frac{k'}{\alpha}\right)\left(\mathbb{I}\left(\pi_{[k]}; Y\right) - \delta\right)$. This gives us $\mathbb{I}\left(\pi_{\mathsf{Greedy}[k']}; Y\right) \geq \left(\mathbb{I}\left(\pi_{[k]}; Y\right) - \delta\right)\left(1 - \exp\left(-\frac{k'}{k\gamma \max\{\log n, \log \frac{1}{\delta}\}}\right)\right)$. ■

5.1. Proof of Lemma 3

We first show a sufficient condition for Lemma 3 as follows.

Lemma 4 *Fix any $\alpha \in [0, \log n]$. If we assume $\max_{e\in E} \mathbb{I}(X_e; Y)$ is sufficiently small, i.e.,*

$$\max_{e\in E} \mathbb{I}(X_e; Y) \leq \frac{\alpha}{k\gamma \max\{\log n, \log \frac{1}{\alpha}\}} \triangleq I_0(\alpha). \tag{11}$$

then no policy of length k is able to achieve a mutual information α, i.e., $\forall \pi_{[k]}$, $\mathbb{I}\left(\pi_{[k]}; Y\right) < \alpha$.

The sufficiency is immediate: suppose Lemma 4 holds. Now, if there exists a policy π of length k with mutual information $\mathbb{I}(\pi; Y)$, then by letting $\alpha = \mathbb{I}\left(\pi_{[k]}; Y\right)$, it must hold that $\max_{e\in[m]} \mathbb{I}(X_e; Y) > \frac{\mathbb{I}\left(\pi_{[k]};Y\right)}{k\gamma \max\{\log n, \log \frac{1}{\mathbb{I}\left(\pi_{[k]};Y\right)}\}}$, which gives us Lemma 3. So for the rest of the proof, we focus on proving Lemma 4. We assume that (11) holds. We consider a policy[5] π of length k and show that $\mathbb{I}(\pi, Y) < \alpha$. We divide the proof into three steps.

Step 1: Recall that each r.v. D_e, as a deterministic function of Y, takes values over its support $\mathcal{D}_e$ with distribution $\Pr(D_e = d_e)$ (and these distributions are possibly different for each $e \in [m]$). Denote $p_{e,\max} = \max_{d_e\in\mathcal{D}_e} \Pr(D_e = d_e)$ to be the probability of the most-likely outcome for D_e. Let us first see how the value $\mathbb{I}(X_e; Y)$ can be expressed in terms of $p_{e,\max}$ and $S(W_e)$. We first argue that for any $e \in [m]$ we have that $\mathbb{I}(X_e; Y) = \mathbb{I}(X_e; D_e)$. This is because D_e is a function of Y, and X_e is a function of noise N (which is independent of Y) and D_e (i.e., $Y \to D_e \to X_e$ forms a Markov chain).

Lemma 5 *Fix $\theta \in (0, 1/4]$. If $\mathbb{I}(X_e; D_e) \leq \theta S(W_e)$, then we have $p_{e,\max} \geq (1 + \sqrt{1 - 4\theta})/2$.*

We relegate the proof of this Lemma to the appendices. By combining Lemma 5 with Inequality (11) and the fact that $\mathbb{I}(X_e; Y) = \mathbb{I}(X_e; D_e)$, we obtain that for any $e \in [m]$ (note the fact that $I_0/S_{\min} \leq \frac{1}{4}$),

$$p_{e,\max} \geq \frac{1}{2}\left(1 + \sqrt{1 - 4\theta}\right) \geq \frac{1}{2}\left(1 + \sqrt{1 - 4\frac{I_0}{S_{\min}}}\right) \triangleq \beta \tag{12}$$

Step 2: This is the situation at level 0. In order to investigate how the values $p_{e,\max}$ change as we perform more tests and observe their outcomes, we have to take into account how the the noise affects the prior and so on. However, we intend to avoid such an analysis. For this purpose, we first prove that the performance of the system would only become better if we were given full information about the noise N. Formally speaking, as mentioned above, a policy π which has length k can also be thought of as a random object which outputs a set of k test-outcome pairs $\psi_\pi \triangleq \{(e_{\pi,1}, x_{e_{\pi,1}}), \ldots, (e_{\pi,k}, x_{e_{\pi,k}})\}$. Such a policy starts from a root node (with only the knowledge about the probabilistic model of Y and X_e's) and performs its tests sequentially and adaptively according to what it has observed. Now, consider an other random variable G defined as follows:

$$G = \{(e_{\pi,1}, d_{e_{\pi,1}}), (e_{\pi,2}, d_{e_{\pi,2}}), \ldots, (e_{\pi,k}, d_{e_{\pi,k}})\}. \tag{13}$$

One can think of G as an *oracle* sitting besides the system π and observing its actions. Furthermore, at each time whatever test e that π picks, G has access to the outcome of D_e, i.e., d_e. Note that G does not know the true value of Y. But we expect that G has a better idea about Y than π has. This is because G knows the deterministic outcomes of the tests that π has picked while π only knows

5. To simplify notation, from now on we use π instead of $\pi_{[k]}$.

a noisy version of these deterministic outcomes (i.e., what π observes is a noisy version of what G observes). Let N be a random vector concatenating all N_e. Indeed, we can write

$$\mathbb{I}(G;Y) = \mathbb{H}(Y) - \mathbb{H}(Y \mid G) \overset{(a)}{=} \mathbb{H}(Y) - \mathbb{H}(Y \mid G, N) \overset{(b)}{\geq} \mathbb{H}(Y) - \mathbb{H}(Y \mid \pi) = \mathbb{I}(\pi;Y), \quad (14)$$

where (a) follows from the fact that Y is independent from N, and (b) is because the output of π is a deterministic function of the the noise N and the output of G. The idea now is to analyze G. Note that: (i) Since G has access to the deterministic values d_e of the tests that π picks, its posterior about Y is decoupled from the noise N. (ii) Any upper bound on $\mathbb{I}(G;Y)$ would also be an upper bound on $\mathbb{I}(\pi;Y)$ by (14).

Step 3: Let us now find an upper bound on $\mathbb{I}(G;Y)$. For this, we start from the root node of π. Recall that in Step 1 we proved that any of the tests X_e satisfies the relation (12). In other words, at time 0 (before performing any tests by π), if we define for $e \in [m]$

$$b_e = \underset{b \in \mathcal{D}}{\operatorname{argmax}} \{\Pr(D_e = b)\}, \text{ and } \mathcal{Y}_e = \{y \in \mathcal{Y} : D_e(y) = b_e\}, \quad (15)$$

then by (12) we have that

$$\Pr(y \in \mathcal{Y}_e^{\complement}) = \Pr(D_e(Y) \neq b_e) = 1 - p_{e,\max} \leq 1 - \beta, \quad (16)$$

where by $\mathcal{Y}_e^{\complement}$ we mean the complement of the set $\mathcal{Y}_e$. The policy π has length k, i.e., it sequentially and adaptively performs tests which we denote by $e_{\pi,1}, \cdots, e_{\pi,k}$ and the choice of $e_{\pi,i}$ is based on the full observation of the outcomes of $e_{\pi,1}, \cdots, e_{\pi,i-1}$. We now consider the following event

$$\mathcal{A} = \{(D_{e_{\pi,1}} = b_{e_{\pi,1}}) \wedge (D_{e_{\pi,2}} = b_{e_{\pi,2}}) \wedge \cdots \wedge (D_{e_{\pi,k}} = b_{e_{\pi,k}})\}, \quad (17)$$

I.e., $\mathcal{A}$ is the event that whatever test e that π picks, its deterministic part D_e outputs its most likely outcome b_e. We establish a lower bound on the probability of $\mathcal{A}$ through the following lemma:

Lemma 6 *If for every test $e \in [m]$ we have $p_{e,\max} \geq \beta$, then $\Pr(\mathcal{A}) \geq 1 - k(1-\beta)$.*

We relegate the proof of this lemma to the appendices. In other words, the random variable G has observations compatible with the event $\mathcal{A}$ with probability at least as the lower bound provided in Lemma 6. We can now write

$$\begin{aligned}\mathbb{H}(Y \mid G) &\overset{(a)}{=} \Pr(\mathcal{A})\mathbb{H}(Y \mid G, \mathcal{A}) + (1 - \Pr(\mathcal{A}))\mathbb{H}\left(Y \mid G, \mathcal{A}^{\complement}\right) \\ &\overset{(b)}{\geq} \Pr(\mathcal{A})\mathbb{H}(Y \mid G, \mathcal{A}). \end{aligned} \quad (18)$$

Here, (a) follows from the fact that the event $\mathcal{A}$ is a function of what G observes. Also, (b) follows from entropy function being positive. It remains to find a lower bound on $\mathbb{H}(Y \mid G, \mathcal{A})$. For this, note that if we end up being in the event $\mathcal{A}$, then all the hypotheses in the set $\cap_{j=1}^{k} \mathcal{Y}_{e_{\pi,j}}$ would remain compatible with the observations that G has had. Let us assume that G has observed $\{e_{\pi,1}, \ldots, e_{\pi,k}\}$ and the corresponding outcomes $\{D_{e_{\pi,1}} = b_{e_{\pi,1}}, \ldots, D_{e_{\pi,k}} = b_{e_{\pi,k}}\}$ (so that event $\mathcal{A}$ has taken place). To simplify notation, let us define $U \triangleq \cap_{j=1}^{k} \mathcal{Y}_{e_{\pi,j}}$. By the union bound and (12) we have

$$\Pr(U) = 1 - \Pr(U^{\complement}) \geq 1 - \sum_{j=1}^{k} (1 - \Pr(\mathcal{Y}_{e_{\pi,j}})) \geq 1 - k(1-\beta). \quad (19)$$

Now, the posterior that G has about Y, $\Pr(Y \mid y \in U)$, would become as follows:

$$\text{If } y \in U, \text{then } \Pr(y \mid y \in U) = \frac{\Pr(y)}{\Pr(U)}, \qquad \text{and} \qquad \text{If } y \notin U, \text{then } \Pr(y \mid y \in U) = 0.$$

The entropy of the posterior then becomes

$$\begin{aligned}\mathbb{H}(Y \mid y \in U) = \sum_{y \in U} \frac{\Pr(y)}{\Pr(U)} \log \frac{\Pr(U)}{\Pr(y)} &= \frac{1}{\Pr(U)} \sum_{y \in X} \Pr(y) \log \frac{1}{\Pr(y)} + \log \Pr(U) \\ &\overset{(a)}{\geq} \frac{\mathbb{H}(Y)}{\Pr(U)} - \frac{1 - \Pr(U)}{\Pr(U)} \log \frac{n}{1 - \Pr(U)} + \log \Pr(U) \\ &\overset{(b)}{\geq} \mathbb{H}(Y) - \frac{1-\rho}{\rho} \log \frac{n}{1-\rho} + \log(\rho). \end{aligned} \tag{20}$$

Here, step (a) follows Lemma 8 as stated in Appendix A. In step (b) we have assumed that $\rho \triangleq 1 - k(1-\beta)$. We thus have from (19) that $\Pr(U) \geq \rho$, and step (b) follows from simple calculus. We also note from Lemma 6 that $\Pr(\mathcal{A}) \geq \rho$. Hence, given event $\mathcal{A}$, the entropy of the posterior that G has about Y is always lower bounded by (20). We thus obtain from (18) that

$$\mathbb{H}(Y \mid G) \geq \Pr(\mathcal{A}) \mathbb{H}(Y \mid G, \mathcal{A}) \geq \rho \mathbb{H}(Y) - (1-\rho) \log \frac{n}{1-\rho} + \rho \log \rho.$$

Finally, we obtain

$$\begin{aligned}\mathbb{I}(G;Y) = \mathbb{H}(Y) - \mathbb{H}(Y \mid G) &\leq \mathbb{H}(Y) - \rho \mathbb{H}(Y) + (1-\rho) \log \frac{n}{1-\rho} + \rho \log \frac{1}{\rho} \\ &= (1-\rho)(\mathbb{H}(Y) + \log n) + (1-\rho) \log \frac{1}{1-\rho} + \rho \log \frac{1}{\rho}. \end{aligned} \tag{21}$$

From now on, we assume that $k \geq 2$ (this is because the result of Lemma 4 is clear when $k = 1$). By the definition I_0 in (21), we have $\frac{I_0}{S_{\min}} = \frac{\alpha}{7k \max\{\log n, \log \frac{1}{\alpha}\}} \leq \frac{1}{14}$ due to the fact that $\alpha \leq \log n$. By the definition of β in (12), we have $1 - \rho = k(1-\beta) = k\left(\frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\frac{I_0}{S_{\min}}}\right) < \frac{11 k I_0}{10 S_{\min}}$ when $\frac{I_0}{S_{\min}} \leq \frac{1}{14}$.[6] Using (11) we get $1 - \rho < \frac{11}{10} \frac{\alpha}{7 \max\{\log n, \log \frac{1}{\alpha}\}} < \frac{11\alpha}{70 \log n}$. Now, from (21) we have

$$\begin{aligned}\mathbb{I}(G;Y) &< \frac{11\alpha}{70} \frac{\mathbb{H}(Y) + \log n}{\log n} + (1-\rho) \log \frac{1}{1-\rho} + (1 - (1-\rho)) \log \frac{1}{1 - (1-\rho)} \\ &\overset{(a)}{<} \frac{11\alpha}{35} + (1-\rho) \left(\log \frac{1}{1-\rho} + \frac{1}{\ln 2} \right) \\ &< \frac{11\alpha}{35} + \frac{11\alpha}{70} \frac{\log \frac{70}{11} + \log\left(\max\{\log n, \log(1/\alpha)\}\right) + \log \frac{1}{\alpha} + \frac{1}{\ln 2}}{\max\{\log n, \log(1/\alpha)\}} \\ &\overset{(b)}{\leq} \frac{11\alpha}{35} + \frac{11\alpha}{70} \left(\frac{\log \frac{70}{11} + \log \log n + \frac{1}{\ln 2}}{\log n} + 1 \right) \overset{(c)}{<} \alpha. \end{aligned}$$

Here, (a) follows from the fact that $\mathbb{H}(Y)$ is less than $\log n$ (because $|\mathcal{Y}| = n$), and the inequality $-(1-x)\log(1-x) < x/(\ln 2)$ for $x \in (0,1)$. Also, (b) follows from simple calculus steps which we omit for the sake of space, and (c) simply follows when $n \geq 3$. For $n = 2$, the proof of Lemma 3 can be done in a simpler way as above and we relegate it to Appendix B.

6. To prove this, one can show that the function $f(x) = \left(1/2 - (1/2)\sqrt{1-4x}\right)/x$ is monotone increasing for $x \in (0, 1/14]$. By plugging in $x = 1/14$ we obtain $f(1/14) \leq 11/10$.

Acknowledgments

This work was supported in part by the DARPA MSEE FA8650-11-1-7156, ERC StG 307036, a Microsoft Research Faculty Fellowship, and a Google European Doctoral Fellowship.

References

Maria-Florina Balcan and Ruth Urner. Active learning–modern learning theory. 2015.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 65–72. ACM, 2006.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *Learning Theory*, pages 35–50. Springer, 2007.

Scott M Berry, Bradley P Carlin, J Jack Lee, and Peter Muller. *Bayesian adaptive methods for clinical trials*. CRC press, 2010.

Marat Valievich Burnashev. Data transmission over a discrete channel with feedback. random transmission time. *Problemy peredachi informatsii*, 12(4):10–30, 1976.

Venkatesan T Chakaravarthy, Vinayaka Pandit, Sambuddha Roy, Pranjal Awasthi, and Mukesh Mohania. Decision trees for entity identification: Approximation algorithms and hardness results. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 53–62. ACM, 2007.

Yuxin Chen, Shervin Javdani, Amin Karbasi, James Andrew Bagnell, Siddhartha Srinivasa, and Andreas Krause. Submodular surrogates for value of information. In *Proc. Conference on Artificial Intelligence (AAAI)*, 2015.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, pages 755–770, 1959.

Imre Csiszar and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

Sanjoy Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems 17*. 2005.

Sanjoy Dasgupta and J Langford. Active learning. *Encyclopedia of Machine Learning*, 2011.

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.

Valerii Vadimovich Fedorov. *Theory of optimal experiments*. Elsevier, 1972.

Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133–168, 1997.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *JAIR*, 2011.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. In *Proc. Neural Information Processing Systems (NIPS)*, December 2010.

Alon Gonen, Sivan Sabato, and Shai Shalev-Shwartz. Efficient active learning of halfspaces: an aggressive approach. *The Journal of Machine Learning Research*, 14(1):2583–2615, 2013.

Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th international conference on Machine learning*, pages 353–360. ACM, 2007.

Steve Hanneke. *Statistical Theory of Active Learning*. Now Publishers Incorporated, 2014.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *CoRR*, abs/1410.0996, 2014.

Michael Horstein. Sequential transmission using noiseless feedback. *Information Theory, IEEE Transactions on*, 9(3):136–143, 1963.

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

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.

Chun-Wa Ko, Jon Lee, and Maurice Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

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.

Andreas Krause and Carlos Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2005.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

SP Luttrell. The use of transinformation in the design of data sampling schemes for inverse problems. *Inverse Problems*, 1(3):199, 1985.

David MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

Mohammad Naghshvar, Tara Javidi, and Kamalika Chaudhuri. Noisy bayesian active learning. *CoRR*, abs/1312.2315, 2013a.

Mohammad Naghshvar, Tara Javidi, et al. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013b.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1):265–294, 1978.

Robert Nowak. Noisy generalized binary search. In *Advances in neural information processing systems*, pages 1366–1374, 2009.

Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948.

Alexandre B Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166, 2004.

Jason L Williams, John Iii, and Alan S Willsky. Performance guarantees for information theoretic active inference. In *International Conference on Artificial Intelligence and Statistics*, pages 620–627, 2007.

Alice X. Zheng, Irina Rish, and Alina Beygelzimer. Efficient test selection in active diagnosis via entropy approximation. In *In Proceedings of UAI-05*, 2005.

Appendix A. Supplemental Proofs

Lemma 7 (Long version of Lemma 5) *Consider random variables U and V where U takes values over the set $[t]$ with distributions p and V takes values over $[t']$ with distribution p'. We think of p (resp. p') as a row vector with size t (resp. t') which sums up to one. Furthermore, assume that $p' = pQ$ where $Q = [q_{i,j}]_{t\times t'}$ is a stochastic matrix and $q_{i,j} = \Pr(V = j \mid U = i)$, for $i \in [t]$ and $j \in [t']$. Let $Q_1, Q_2, \cdots, Q_t$ denote the rows of Q and define the minimum distance of Q as*

$$S = \Big(\min_{i,j\in[t]:i\neq j} |Q_i - Q_j|_{\mathrm{TV}}\Big)^2.$$

Also, define $p_{\max} = \max_{i\in[t]} p_i$ and $u_{\max} = \max_{i\in[t]} p_i(1-p_i)$. Then, we have

1. $\mathbb{I}(U;V) \geq S u_{\max}$.
2. $\mathbb{I}(U;V) \geq \frac{1}{2} S(1 - p_{\max})$.
3. *If $\mathbb{I}(U;V) \leq \theta S$, then we have $p_{\max} \geq 1 - 2\theta$. Also, if $\theta \leq \frac{1}{4}$, then $p_{\max} \geq \frac{1+\sqrt{1-4\theta}}{2}$.*

Proof Let $p = [p_i]$ and $p' = [p'_i]$. Using the introduced notation, we have $\Pr(U = i, V = j) = p_i q_{i,j}$. We thus can write

$$\begin{aligned}\mathbb{I}(U;V) &= \sum_{i=1}^{t}\sum_{j=1}^{t'} p_i q_{i,j} \log \frac{p_i q_{i,j}}{p_i p'_j} \\ &= \sum_{i=1}^{t} p_i \sum_{j=1}^{t'} q_{i,j} \log \frac{q_{i,j}}{p'_j}\end{aligned}$$

$$= \sum_{i=1}^{t} p_i D_{\mathrm{KL}}(Q_i || p')$$
$$\geq 2 \sum_{i=1}^{t} p_i |Q_i - p'|_{\mathrm{TV}}^2, \tag{22}$$

where the last step is due to Pinsker's inequality (Csiszar and Körner, 2011, p 44). For any $j \in [t]$ we can write

$$\begin{aligned}
\sum_{i=1}^{t} p_i |Q_i - p'|_{\mathrm{TV}}^2 &= p_j |Q_j - p'|_{\mathrm{TV}}^2 + \sum_{i:i\neq j} p_i |Q_i - p'|_{\mathrm{TV}}^2 \\
&\geq (1-p_j) p_j |Q_j - p'|_{\mathrm{TV}}^2 + p_j \sum_{i:i\neq j} p_i |Q_i - p'|_{\mathrm{TV}}^2 \\
&= \sum_{i:i\neq j} p_j p_i |Q_j - p'|_{\mathrm{TV}}^2 + \sum_{i:i\neq j} p_j p_i |Q_i - p'|_{\mathrm{TV}}^2 \\
&= \sum_{i:i\neq j} p_j p_i (|Q_j - p'|_{\mathrm{TV}}^2 + |Q_i - p'|_{\mathrm{TV}}^2) \\
&\geq \sum_{i:i\neq j} p_j p_i \frac{(|Q_j - p'|_{\mathrm{TV}} + |Q_i - p'|_{\mathrm{TV}})^2}{2} \\
&\overset{(a)}{\geq} \sum_{i:i\neq j} p_j p_i \frac{|Q_j - Q_j|_{\mathrm{TV}}^2}{2} \\
&\geq \sum_{i:i\neq j} p_j p_i \frac{S}{2} \\
&= p_j (1-p_j) \frac{S}{2}.
\end{aligned} \tag{23}$$

Here, step (a) follows from the triangular inequality for total variation distances. The proof of part (1) is then complete by combining (22) and (23).

For the second part of the lemma, assume w.l.o.g that $t = 2s+1$ and $p_1 \geq p_2 \geq \cdots \geq p_t$ (if t is even we can always let $t \leftarrow t+1$ and let $p_t = 0$). We can then write

$$\begin{aligned}
\sum_{i=1}^{t} p_i |Q_i - p'|_{\mathrm{TV}}^2 &\geq \sum_{j=1}^{s} \left\{ p_{2j-1} |Q_{2_{j-1}} - p'|_{\mathrm{TV}}^2 + p_{2j} |Q_{2j} - p'|_{\mathrm{TV}}^2 \right\} \\
&\geq \sum_{j=1}^{s} p_{2j} \left(|Q_{2_{j-1}} - p'|_{\mathrm{TV}}^2 + |Q_{2j} - p'|_{\mathrm{TV}}^2 \right) \\
&\geq \sum_{j=1}^{s} p_{2j} \frac{\left(|Q_{2_{j-1}} - p'|_{\mathrm{TV}} + |Q_{2j} - p'|_{\mathrm{TV}} \right)^2}{2} \\
&\overset{(a)}{\geq} \sum_{j=1}^{s} p_{2j} \frac{|Q_{2_{j-1}} - Q_{2j}|_{\mathrm{TV}}^2}{2}
\end{aligned}$$

$$\geq \frac{S}{2}\sum_{j=1}^{s} p_{2j}$$
$$\overset{(b)}{\geq} \frac{S}{4}(1-p_1).$$

Here, step (a) follows from the triangle inequality for total variation distances and (b) follows from the fact that when p_i's are decreasing we have $\sum_{j=1}^{s} p_{2j} \geq \frac{1-p_1}{2}$. Part 2 is now proven by using the above derivation and (22).

Part 3 simply follows from the fact that in the assumption $\mathbb{I}(U;V) \leq \frac{\theta S}{4}$ holds, then by part 2 we have $p_{\max} \geq 1-2\theta$. Also, if $\theta \leq 1/4$, then $p_{\max} \geq 1/2$ and from part 1 we get $p_{\max}(1-p_{\max}) \leq \theta$ and putting the two together we get the result. ■

Proof [Proof of Lemma 6] Let us illustrate the idea by first assuming that π has length two (see

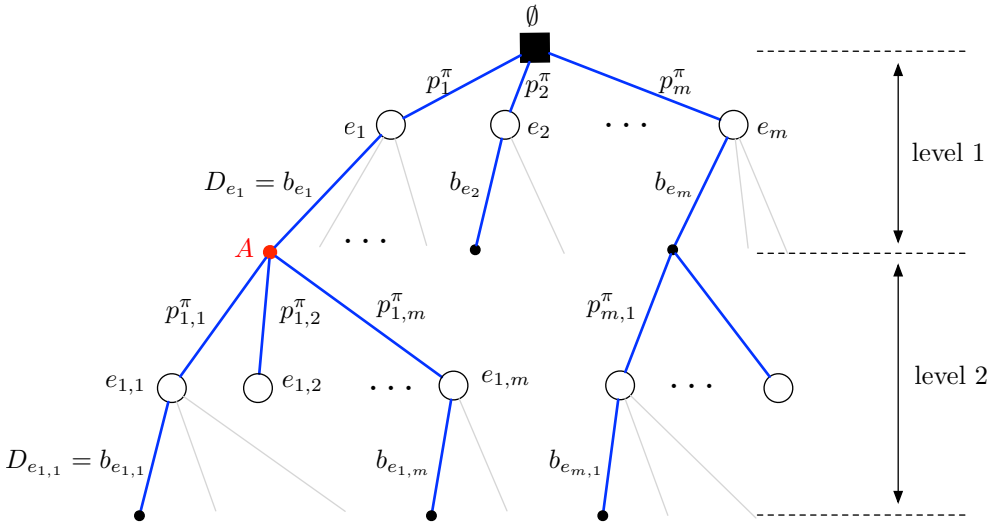


Figure 4: A two-stage decision tree representation for (stochastic) policy $\pi_{(2)}$

Figure 4). Afterwards, we prove the statement for π with an arbitrary length k. Recall that we consider randomised policies too. In the very beginning, when no observations have been made, π can possibly choose any of the possible m tests $e_1, \cdots, e_m \in [m]$. We thus assume that π chooses e_i with probability p_i^π. Furthermore, the choice of e_i clearly does not reveal any information about Y (because we only talk about the choice and hence no observations have been done so far). Let us define

$$\mathcal{A}_r = \{(D_{e_1} = b_{e_1}) \wedge (D_{e_2} = b_{e_2}) \wedge \cdots \wedge (D_{e_r} = b_{e_r})\},$$

to be the event that the deterministic part of the first r tests that π has picked all have the most-likely outcome. Once π chooses its first test (let's say e_1), the event $\mathcal{A}_1$ takes place only if the output of D_{e_1} is precisely equal to b_{e_1}. Hence, in Figure 4, in order to find $\Pr(\mathcal{A}_1)$ we need to add up the probabilities of the blue paths up to level 1. As a result,

$$\Pr(\mathcal{A}_1) = \sum_{i=1}^{m} p_i^\pi \Pr(D_{e_i} = b_{e_i})$$

$$\geq \sum_{i=1}^{m} p_i^{\pi} \beta = \beta.$$

Now, let us see what happens when π selects its second test. For this, assume for simplicity that the first choice of π was e_1 and the output of D_{e_1} is indeed b_{e_1} (i.e., we are standing at point A on the tree depicted in Figure 4). At this moment, the noise affects the deterministic outcome of e_1 (which we have assumed to be b_{e_1}) and hence π observes a noisy version of b_{e_1}. Based on this observation, π selects a new test (which might be a randomised selection). Let us assume that this time π selects the i-th test with probability $p_{1,i}^{\pi}$. An important point to note here is that conditioned on the fact that $D_{e_1} = b_{e_1}$ (i.e. point A on the tree), the new *choice* of π does not give any new information about Y. This is because conditioned on $D_{e_1} = b_{e_1}$, the choice of π is only a function of b_{e_1} and the noise and possibly some other random variables (used to randomise the policy) that are independent of Y given $D_{e_1} = b_{e_1}$. Hence, we can write

$$\Pr(\mathcal{A}_2 \mid D_{e_1} = b_{e_1}) = \sum_{i=1}^{m} p_{1,i}^{\pi} \Pr(D_{e_i} = b_{e_i} \mid D_{e_1} = b_{e_1})$$

Indeed, the above argument is valid if π had chosen any generic test e_i (instead of e_1) as its first test. The value $\Pr(\mathcal{A}_2)$ can now be found by summing up the probabilities of all the points at level 2 that are the end-point of a blue path. We have

$$\begin{aligned}
\Pr(\mathcal{A}_2) &= \sum_{i=1}^{m} p_i^{\pi} \Pr(D_{e_i} = b_{e_i}) \Pr(\mathcal{A}_2 \mid D_{e_i} = b_{e_i}) \\
&= \sum_{i,j=1}^{m} p_i^{\pi} p_{i,j}^{\pi} \Pr(D_{e_i} = b_{e_i}) \Pr(D_{e_j} = b_{e_j} \mid D_{e_i} = b_{e_i}) \\
&= \sum_{i,j=1}^{m} p_i^{\pi} p_{i,j}^{\pi} \Pr(D_{e_j} = b_{e_j}, D_{e_i} = b_{e_i}) \\
&\overset{(a)}{\geq} \sum_{i,j=1}^{m} p_i^{\pi} p_{i,j}^{\pi} (1 - 2(1-\beta)) \\
&\geq (1 - 2(1-\beta)) \sum_{i=1}^{m} p_i^{\pi} \sum_{j=1}^{m} p_{i,j}^{\pi} \\
&= 1 - 2(1-\beta),
\end{aligned}$$

where (a) follows from the Union bound.

Now consider the general case where π has length k. As explained before, the event $\mathcal{A}$ happens only on the "good paths" (i.e., the paths that event $\mathcal{A}$ happens, as depicted in blue in Figure 5) of the policy tree. Define path $\psi_i := \left\{(e_{\pi,1}, D_{e_{\pi,1}} = b_{e_{\pi,1}}), \ldots, (e_{\pi,i}, D_{e_{\pi,i}} = b_{e_{\pi,i}})\right\}$. We then have

$$\begin{aligned}
\Pr(\mathcal{A}_k) &= \sum_{\psi_k} \Pr\left((e_{\pi,1}, D_{e_{\pi,1}} = b_{e_{\pi,1}}), \ldots, (e_{\pi,k}, D_{e_{\pi,k}} = b_{e_{\pi,k}})\right) \\
&= \sum_{\psi_k} \Pr\left(e_{\pi,1}\right) \Pr\left(D_{e_{\pi,1}} = b_{e_{\pi,1}}\right) \prod_{i=1}^{k-1} \Pr(e_{\pi,i+1} \mid \psi_i) \times
\end{aligned}$$

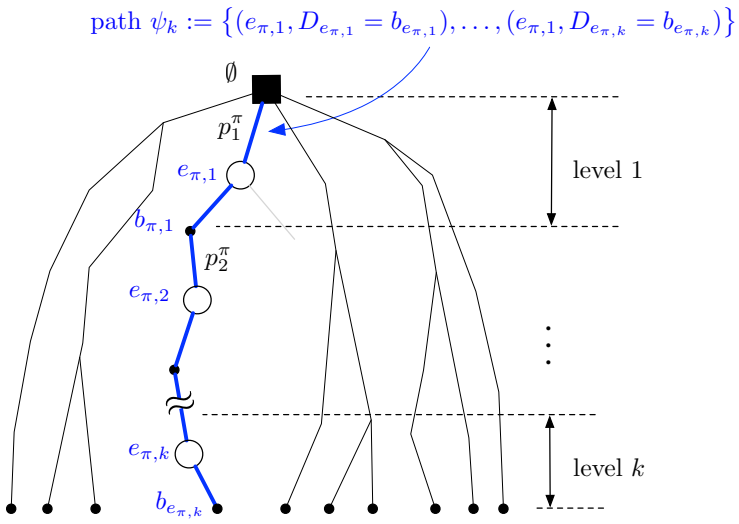


Figure 5: Event $\mathcal{A}$ in the policy tree

$$\prod_{i=1}^{k-1} \Pr\left(D_{e_{\pi,i+1}} = b_{e_{\pi,i+1}} \mid D_{e_{\pi,1}} = b_{e_{\pi,1}}, \ldots, D_{e_{\pi,i}} = b_{e_{\pi,i}}\right)$$

$$= \sum_{\psi_k} \Pr\left(D_{e_{\pi,1}} = b_{e_{\pi,1}}, \ldots, D_{e_{\pi,k}} = b_{e_{\pi,k}}\right) \times \Pr\left(e_{\pi,1}\right) \prod_{i=1}^{k-1} \Pr\left(e_{\pi,i+1} \mid \psi_i\right)$$

Since for each $e_{\pi,i}$ it holds that $\Pr(D_{e_{\pi,i}} = b_{e_{\pi,i}}) \geq \beta$, applying the union bound we obtain

$$\Pr\left(D_{e_{\pi,1}} = b_{e_{\pi,1}}, \ldots, D_{e_{\pi,k}} = b_{e_{\pi,k}}\right) \geq 1 - k(1-\beta).$$

Thus,

$$\begin{aligned}
\Pr(\mathcal{A}_k) &\geq (1 - k(1-\beta)) \sum_{\psi_k} \Pr\left(e_{\pi,1}\right) \prod_{i=1}^{k-1} \Pr\left(e_{\pi,i+1} \mid \psi_i\right) \\
&= (1 - k(1-\beta)) \sum_{\psi_{k-1}} \Pr\left(e_{\pi,1}\right) \prod_{i=1}^{k-2} \Pr\left(e_{\pi,i+1} \mid \psi_i\right) \sum_{e \in [m]} \Pr\left(e_{\pi,k} = e \mid \psi_{k-1}\right) \\
&= (1 - k(1-\beta)) \sum_{\psi_{k-2}} \Pr\left(e_{\pi,1}\right) \prod_{i=1}^{k-3} \Pr\left(e_{\pi,i+1} \mid \psi_i\right) \sum_{e \in [m]} \Pr\left(e_{\pi,k-1} = e \mid \psi_{k-2}\right) \\
&\quad \vdots \\
&= (1 - k(1-\beta)) \sum_{\psi_2} \Pr\left(e_{\pi,1}\right) \Pr\left(e_{\pi,2} \mid \psi_1\right) \sum_{e \in [m]} \Pr\left(e_{\pi,3} = e \mid \psi_2\right) \\
&= (1 - k(1-\beta)) \sum_{\psi_1} \Pr\left(e_{\pi,1}\right) \sum_{e \in [m]} \Pr\left(e_{\pi,2} = e \mid \psi_1\right)
\end{aligned}$$

$$= (1 - k(1-\beta))$$

■

Lemma 8 *Consider a distribution $p(\cdot)$ a set $\mathcal{Y}$ with $|\mathcal{Y}| = n$. For a subset $X \subseteq Y$ we have*

$$\sum_{y \in X} p_y \log \frac{1}{p_y} \geq \mathbb{H}(p) - (1 - p(X)) \log n + (1 - p(X)) \log(1 - p(X)). \tag{24}$$

Proof We have

$$\begin{aligned}
\mathbb{H}(p) - \sum_{y \in X} p_y \log \frac{1}{p_y} &= \sum_{y \in X^c} p_y \log \frac{1}{p_y} \\
&= p(X^c) \sum_{y \in X^c} \frac{p_y}{p(X^c)} \log \frac{p(X^c)}{p_y} - p(X^{\complement}) \log p(X^{\complement}) \\
&\overset{(a)}{\leq} p(X^{\complement}) \log n - p(X^{\complement}) \log p(X^{\complement}) \\
&= (1 - p(X)) \log n - (1 - p(X)) \log(1 - p(X)),
\end{aligned}$$

where step (a) is due to the fact that the cardinality of the set $X^{\complement}$ is at most n and thus the entropy of any distribution on this set is less than $\log n$. ■

Appendix B. Proof of Lemma 3 for $n = 2$

For $n = 2$ we have $Y = \text{Bernoulli}(p)$. Assume w.l.o.g that $p \leq 1/2$. Each D_e is a deterministic function of Y. So D_e is itself a binary random variable. Now, there exists $e' \in [m]$ such that $\mathbb{I}(D_{e'}; Y) > 0$, otherwise any policy gains zero mutual information and the result of Lemma 3 is trivial. We assume w.l.o.g that $D_{e'} = Y$. By using part 2 of Lemma 7 we get that $\mathbb{I}(X_{e'}; Y) \geq p\frac{S_{\min}}{2}$. Note that $H(Y) = h_2(p)$, where $h_2(x) \triangleq -x \log x - (1-x)\log(1-x)$. Also, it is easy to verify that for $p \leq 1/2$ we have $h_2(p) \leq -2p \log p$ and also $-\log(h_2(p)) \geq -\frac{\log p}{3}$. We thus get that

$$\mathbb{I}(X_{e'}; Y) \geq \frac{S_{\min}}{12} \frac{h_2(p)}{\log 1/h_2(p)}. \tag{25}$$

Now, note that any policy can have at most $\mathbb{I}(\pi, Y) \leq H(Y) = h_2(p)$. Thus, $\log 1/\mathbb{I}(\pi, Y) \geq \log 1/h_2(p)$. As a result,

$$\frac{\mathbb{I}(\pi, Y)}{\log 1/\mathbb{I}(\pi, Y)} \leq \frac{h_2(p)}{\log 1/h_2(p)}. \tag{26}$$

To get the result of Lemma 3, we assume two cases: (i) $\mathbb{I}(\pi, Y) \leq \frac{1}{2}$: in this case $\log 1/\mathbb{I}(\pi, Y) \geq 1$ and by (25) and (26) we obtain that $\mathbb{I}(X_{e'}; Y) \geq \frac{S_{\min}\mathbb{I}(\pi,Y)}{12 \log 1/\mathbb{I}(\pi,Y)}$ (ii) $\mathbb{I}(\pi, Y) > \frac{1}{2}$ which, by using $h_2(p) > \frac{1}{2}$, means that $p > 0.1102$ and thus $\frac{p}{h_2(p)} \geq 1/6$. In this case, we have $\mathbb{I}(X_{e'}; Y) \geq p\frac{S_{\min}}{2} \geq h_2(p) S_{\min} \frac{p}{2h_2(p)} \geq \frac{S_{\min} h_2(p)}{12} \geq \frac{S_{\min}\mathbb{I}(\pi;Y)}{12}$. Thus, form the two cases, we have proven that $\mathbb{I}(X_{e'}; Y) \geq \frac{S_{\min}\mathbb{I}(\pi;Y)}{12 \max\{\log n, \log 1/\mathbb{I}(\pi,Y)\}}$ for any policy π. This proves Lemma 3 for $k \geq 2$. Note that the result of Lemma 3 is trivially valid when $k = 1$.

Appendix C. An example where $S_{\min}$ is necessary

In this section, we introduce an example to show an upper bound on the ratio between the greedy policy and the optimal policy, which involves $S_{\min}$. We first describe the high-level intuition behind the example in Section C.1, and then present details in Section C.2.

C.1. The construction strategy

In our example, we first design $T+1$ sets of tests $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_{T+1}$ (the value of T will be specified later). Tests in these sets have low information gain on their own, as in the beginning, all these tests have 0 outcome with high probability. However, if one test $X^{(i)} \in \mathcal{X}_i$ from set i is observed to have positive outcome, then one can always find a test $X^{(i+1)} \in \mathcal{X}_{i+1}$, that is very informative about the remaining hypotheses that are consistent with the outcome of $X^{(i)}$. A smart policy will (sequentially) pick tests among these sets, and one can show that with at most $2T$ tests, this policy will reduce $H(Y)$ by at least $(\log n)/T$ bits.

To "confuse" the greedy policy, we design another set of tests $\mathcal{Z}$, with infinitely many identical tests, and the deterministic outcome of each test takes value among $\{0, 1\}$ with equal probability. As in Example 1, we assume that observed outcome of each test $Z \in \mathcal{Z}$ is perturbed by a binary symmetric channel with flip over probability ϵ. If $S_{\min} = (1-2\epsilon)^2$ is sufficiently large (in fact, in our example, $S_{\min} \geq \Omega\left(1/\log n\right)$), one can show that with high probability, the greedy policy always picks among this set of tests, and within $2T$ tests, the gain of the greedy policy is at most $O(S_{\min}T)$ Bits. Setting $T = 2\sqrt{\log n}$, the ratio between the gain of greedy and the smarter policy is at most $c_0 S_{\min}$, where c_0 is some constant (note that when $S_{\min}$ is small we have $c_0 S_{\min} \approx (1 - \exp\left(c_0 S_{\min}\right))$).

C.2. Details of the example

Consider the following *treasure hunt* example. Assume that the hidden variable Y takes value among set $\mathcal{Y} = \{y_1, \cdots, y_n\}$ with uniform distribution. Define $T \triangleq 2\sqrt{\log n}$, and let $k' = k = 2T$, i.e., the greedy policy π_{Greedy} and the optimal policy π_{OPT} have the same budget. We hereby design a problem with $\frac{2(10+2\log\log n)}{\log n} \leq S_{\min} \leq \frac{1}{256\sqrt{\log n}(\log\log n)^2}$, such that π_{Greedy} performs considerably worse w.r.t. π_{OPT} (indeed, the ratio is a factor of $S_{\min}$). Finally, note that we choose n sufficiently large so that the bounds provided are meaningful.

We first define $T+1$ types of tests, namely Type 1, Type 2, ..., Type $T+1$. The first type of tests contains a total number of T tests, all of which have binary outcomes. We denote them by $\mathcal{X}_1 = \{X_1^{(1)}, X_2^{(1)}, \ldots X_T^{(1)}\}$. We partition the set $\mathcal{Y}$ into T equal-size groups, denoted by $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_T$, each containing n/T values (assume n is such that this partition is possible). Each test of Type 1 can be thought to be informative about only a small fraction of $\mathcal{Y}_1$. Specifically, imagine that we further partition the set $\mathcal{Y}_1$ into T groups of equal size, denoted by $\mathcal{Y}_{1,1}, \mathcal{Y}_{1,2}, \cdots, \mathcal{Y}_{1,T}$. For each $i \in \{1, \ldots, T\}$, we define

$$X_i^{(1)} = \begin{cases} 1, & \text{if } y \in \mathcal{Y}_{1,i}; \\ 0, & \text{o.w.} \end{cases}$$

The mutual information between each test $X_i^{(1)}$ and Y is

$$\mathbb{I}\left(X_i^{(1)}; Y\right) = \mathbb{H}\left(X_i^{(1)}\right) - \mathbb{H}\left(X_i^{(1)} \mid Y\right)$$

$$= \frac{1}{T^2}\log T^2 + \left(1 - \frac{1}{T^2}\right)\log\left(\frac{T^2}{T^2-1}\right)$$
$$\leq \frac{\log T^2}{T^2} + \frac{2}{T^2}$$
$$= \frac{4 + \log\log n}{4\log n}. \tag{27}$$

The second set of tests also contains T tests, denoted by $\mathcal{X}_2 = \{X_1^{(2)}, X_2^{(2)}, \ldots, X_T^{(2)}\}$. Each test in $\mathcal{X}_2$ has three outcomes. When $y \in \mathcal{Y}_{1,i}$, $X_i^{(2)}$ takes values among $\{1, 2\}$ with equal probability $1/2$; when $y \notin \mathcal{Y}_{1,i}$, $X_i^{(2)} = 0$. For each $i \in \{1, \ldots, T\}$, we denote $\mathcal{Y}_{1,i,1}$ and $\mathcal{Y}_{1,i,2}$ to be the set of values of y on which $X_i^{(2)} = 1$ and $X_i^{(2)} = 2$, then

$$X_i^{(2)} = \begin{cases} 2, & \text{if } y \in \mathcal{Y}_{1,i,2}; \\ 1, & \text{if } y \in \mathcal{Y}_{1,i,1}; \\ 0, & \text{o.w.} \end{cases}$$

The mutual information between each test $X_i^{(2)}$ and Y is

$$\mathbb{I}\left(X_i^{(2)}; Y\right) = \mathbb{H}\left(X_i^{(2)}\right) - \mathbb{H}\left(X_i^{(2)} \mid Y\right)$$
$$= \frac{1}{2T^2}\log 2T^2 + \frac{1}{2T^2}\log 2T^2 + \left(1 - \frac{1}{T^2}\right)\log\left(\frac{T^2}{T^2-1}\right)$$
$$\leq \frac{\log 2T^2}{T^2} + \frac{2}{T^2}$$
$$= \frac{5 + \log\log n}{4\log n}. \tag{28}$$

Further, there are a total number of $2T$ tests of Type 3, i.e., $\mathcal{X}_3 = \{X_1^{(3)}, X_2^{(3)}, \ldots X_{2T}^{(3)}\}$. Each of the tests has 5 outcomes. Intuitively, we design these tests to further refine the set of values Y can take based on the outcome of tests in $\mathcal{X}_2$: if one of the tests in $\mathcal{X}_2$ has non-zero realization, then there exists a test $X_i^{(3)} \in \mathcal{X}_3$ that will help us identify a much smaller subset of $\mathcal{Y}$. Formally, for $i \in \{1, \ldots, T\}, j \in \{1, 2\}$, and $l \in \{1, 2, 3, 4\}$, we denote $\mathcal{Y}_{1,i,j,l}$ to be the set of values of y on which $X_{2i+1-j}^{(3)} = l$, and each $\mathcal{Y}_{1,i,j,l}$ contains $\frac{n}{4\times 2T^2}$ values. We define

$$X_{2i+1-j}^{(3)} = \begin{cases} 4, & \text{if } y \in \mathcal{Y}_{1,i,j,4}; \\ 3, & \text{if } y \in \mathcal{Y}_{1,i,j,3}; \\ 2, & \text{if } y \in \mathcal{Y}_{1,i,j,2}; \\ 1, & \text{if } y \in \mathcal{Y}_{1,i,j,1}; \\ 0, & \text{o.w.} \end{cases}$$

For $i \in \{1, \ldots, 2T\}$, the mutual information between each test $X_i^{(3)}$ and Y is

$$\mathbb{I}\left(X_i^{(3)}; Y\right) = 4 \times \frac{1}{4 \times 2T^2}\log\left(4 \times 2T^2\right) + \left(1 - \frac{1}{2T^2}\right)\log\left(\frac{2T^2}{2T^2-1}\right)$$
$$\leq \frac{\log(4 \times 2T^2)}{2T^2} + \frac{2}{2T^2}$$

$$= \frac{7 + \log\log n}{8 \log n}. \tag{29}$$

Similarly with the above construction, we define tests of Type t, $t \in \{2, \ldots, T+1\}$ to be $\mathcal{X}_t$, with $|\mathcal{X}_t| = \prod_{i=1}^{t-2} 2^i$. Those tests, if sequentially performed, behave as follows: If one of the tests in $\mathcal{X}_{t-1}$ has non-zero realization, then one can perform a test in $X_i^{(t)} \in \mathcal{X}_t$, and the outcome of this test can reduce the number of consistent hypotheses to a factor of $\frac{1}{2^{t-1}}$.

Suppose there is a "smart" policy, denoted by π^s, which works as follows. It first performs all the T tests in $\mathcal{X}_1$, and the probability that one of them has non-zero outcome is $1/T$. If this happens, then π^s sequentially picks T more tests from each of the sets $\mathcal{X}_2, \mathcal{X}_3, \ldots, \mathcal{X}_{T+1}$. Test $X \in \mathcal{X}_i$ will reduce the number of valid values of $\mathcal{Y}$ by a factor of $\frac{1}{2^{i-1}}$. Hence, by noting that,

$$\frac{n}{T^2}\left(\frac{1}{2} \times \frac{1}{4} \times \cdots \times \frac{1}{2^{T+1}}\right) < 1$$

we can see that, if $y \in \mathcal{Y}_1$, then after $2T$ tests, we get the right hypothesis (i.e., the policy π^s reduces $\mathbb{H}(Y \mid \pi)$ to 0). Since $y \in \mathcal{Y}_1$ occurs with probability $1/T$, we can bound the gain of π^s by $\mathbb{I}(\pi^s; Y) \geq (\log n)/T$.

The Greedy Policy We now define another set of tests, and show that with high probability, the greedy policy prefers this set, but performing tests in this set gives a relatively low gain in terms of entropy reduction. Denote this set of tests by $\mathcal{Z}$. There are infinitely many identical tests in $\mathcal{Z}$. For each test $Z_i \in \mathcal{Z}$, the observed outcome is flipped from the (deterministic) outcome D_i given Y with probability ϵ (so that $S_{\min} = (1-2\epsilon)^2$), and the flipping events of the tests are independent (i.e., each test is associated with a binary symmetric noise channel). Assume that initially, the deterministic outcome D_i of Z_i is uniformly distributed among $\{0, 1\}$. In particular, let $\mathcal{Y}^1 \triangleq \mathcal{Y}_1 \cup \cdots \cup \mathcal{Y}_{T/2}$, and $\mathcal{Y}^0 \triangleq \mathcal{Y}_{T/2+1} \cup \cdots \cup \mathcal{Y}_T$, we define for each i,

$$D_i = \begin{cases} 1, & \text{if } y \in \mathcal{Y}^1; \\ 0, & \text{if } y \in \mathcal{Y}^0; \end{cases}$$

and the observed outcome $Z_i = D_i \oplus N_i$, with $Pr(N_i = 1) = \epsilon$.

Then, it is easy to check that in the very beginning (where Y has a uniform distribution) we have $\mathbb{I}(Z_i; Y) = 1 - h_2(\epsilon)$. We prove that the greedy policy π_{Greedy} picks these tests with high probability. The following lemma characterizes such behavior of π_{Greedy}.

Lemma 9 *Assume that* $\frac{2(10+2\log\log n)}{\log n} \leq S_{\min} \leq \frac{1}{256\sqrt{\log n}(\log\log n)^2}$. *With probability at least* $1 - 4\sqrt{\log n}\exp\left(-2(\log\log n)^2\right)$, $\pi_{\mathsf{Greedy}[2T]}$ *will pick* $2T$ *tests in* $\mathcal{Z}$.

The proof of this lemma will appear shortly. Now, note that $\mathbb{H}(Y)$ can at most be $\log n$ under any distribution. So the gain of π_{Greedy} can be bounded from above as follows.

$$\mathbb{I}\left(\pi_{\mathsf{Greedy}}; Y\right) \leq \mathbb{I}(Z_1, \ldots, Z_{2T}; Y) + \left(4\sqrt{\log n}\exp\left(-2(\log\log n)^2\right)\right)\log n.$$

Let us now bound the mutual information term. We have

$$\mathbb{I}(Z_1, Z_2 \ldots, Z_{2T}; Y) = \mathbb{H}(Z_1, Z_2 \ldots, Z_{2T}) - \mathbb{H}(Z_1, Z_2 \ldots, Z_{2T} \mid Y)$$

$$\begin{aligned}
&\leq \sum_{i=1}^{2T} \left(\mathbb{H}\left(Z_i\right) - \mathbb{H}\left(Z_i \mid Y\right)\right) \\
&= \sum_{i=1}^{2T} \mathbb{I}\left(Z_i; D_i\right) \\
&= 2T(1 - h_2(\epsilon))
\end{aligned}$$

As a result, we can write

$$\begin{aligned}
\mathbb{I}\left(\pi_{\mathsf{Greedy}}; Y\right) &\leq 2T(1 - h_2(\epsilon)) + 4\sqrt{\log n} \exp\left(-2(\log\log n)^2\right) \log n \\
&\leq 4TS_{\min} + 4\sqrt{\log n} \exp\left(-2(\log\log n)^2\right) \log n
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{\mathbb{I}\left(\pi_{\mathsf{Greedy}}; Y\right)}{\mathbb{I}\left(\pi^s; Y\right)} &\leq \frac{4TS_{\min} + 4\sqrt{\log n} \exp\left(-2(\log\log n)^2\right) \log n}{\frac{\log n}{T}} \\
&= 16S_{\min} + 8\log n e^{\left(-2(\log\log n)^2\right)} \\
&< 32S_{\min}.
\end{aligned}$$

Hence, the gain of the greedy algorithm (when allowed to choose $2T$ tests) can be at most a fraction $32S_{\min}$ of the optimal algorithm which is allowed to choose also $2T$ tests. It remains to prove Lemma 9.

Proof [Proof of Lemma 9.] The proof goes through the following steps:

Step 1. Consider 2T tests $Z_1, Z_2, \ldots Z_{2T}$ from set $\mathcal{Z}$. If $S_{\min} = (1-2\epsilon)^2 \leq \frac{1}{256\sqrt{\log n}(\log\log n)^2}$, for all $m \leq 2T$, we claim

$$\Pr\left[\frac{1}{e} \leq \left(\frac{1-\epsilon}{\epsilon}\right)^{m-2\sum_{i=1}^m z_i} \leq e\right] \geq 1 - 2\exp\left(-2(\log\log n)^2\right) \tag{30}$$

In the following, we prove the above inequality. By Hoeffding's inequality, we have

$$\Pr\left(\left|\sum_{i=1}^m (1-2z_i) - m(1-2\epsilon)\right| \geq \sqrt{m}\log\log n\right) \leq 2\exp\left(-2(\log\log n)^2\right)$$

Therefore, with probability at least $1 - 2\exp\left(-2(\log\log n)^2\right)$, we have

$$\begin{aligned}
\left(\frac{1-\epsilon}{\epsilon}\right)^{m-2\sum_{i=1}^m z_i} &\leq \left(\frac{1-\epsilon}{\epsilon}\right)^{m(1-2\epsilon)+\sqrt{m}\log\log n} \\
&= e^{\left(\ln\frac{1-\epsilon}{\epsilon}\cdot m\cdot(1-2\epsilon) + \sqrt{m}\ln\frac{1-\epsilon}{\epsilon}\cdot\log\log n\right)} \\
&\leq e^{\left(4(1-2\epsilon)^2 m + 4\sqrt{m(1-2\epsilon)^2}\log\log n\right)} \\
&\leq e^{\left(4\times 2TS_{\min} + 4\sqrt{2TS_{\min}}\log\log n\right)}
\end{aligned}$$

In order for inequality $\left(\frac{1-\epsilon}{\epsilon}\right)^{m-2\sum_{i=1}^m z_i} \le e$ to hold, it suffices to ensure that

$$\begin{cases} 8TS_{\min} \le \frac{1}{2} \\ 4\sqrt{2TS_{\min}}\log\log n \le \frac{1}{2} \end{cases}$$

From the first inequality we get $S_{\min} \le \frac{1}{32\sqrt{\log n}}$; from the second we get $S_{\min} \le \frac{1}{256\sqrt{\log n}(\log\log n)^2}$.

To show $\left(\frac{1-\epsilon}{\epsilon}\right)^{m-2\sum_{i=1}^m z_i} \ge \frac{1}{e}$, we use

$$\begin{aligned}\left(\frac{1-\epsilon}{\epsilon}\right)^{m-2\sum_{i=1}^m z_i} &\ge \left(\frac{1-\epsilon}{\epsilon}\right)^{m(1-2\epsilon)-\sqrt{m}\log\log n} \\ &= e^{\left(\ln\frac{1-\epsilon}{\epsilon}\cdot m\cdot(1-2\epsilon)-\sqrt{m}\ln\frac{1-\epsilon}{\epsilon}\cdot\log\log n\right)},\end{aligned}$$

and it suffices to ensure that $\sqrt{m}\ln\frac{1-\epsilon}{\epsilon}\cdot\log\log n \le 1$, which clearly holds when $S_{\min} \le \frac{1}{256\sqrt{\log n}(\log\log n)^2}$. From (30) and the union bound we get that

$$\Pr\left[\forall m \le 2T : \frac{1}{e} \le \left(\frac{1-\epsilon}{\epsilon}\right)^{m-2\sum_{i=1}^m z_i} \le e\right] \ge 1 - 4\sqrt{\log n}\exp\left(-2(\log\log n)^2\right)$$

Step 2. We now prove Lemma 9 by induction. Assume that $S_{\min} \ge \frac{2(10+2\log\log n)}{\log n}$. By equations (27),(28),(29), we know that the gain of any tests in $\mathcal{X}_t, t \in \{1,\dots,T+1\}$ is less that $S_{\min}$. In the very beginning, $\mathbb{I}(Z_i;Y) = (1-h_2(\epsilon)) \ge S_{\min}$, so $\pi_{\mathsf{Greedy}[2T]}$ chooses a test from $\mathcal{Z}$.

Step 3. Consider an integer $m \le 2T$ and assume that greedy has so far picked tests $Z_1,\dots,Z_m \in \mathcal{Z}$ with outputs $z_1,\dots,z_m$ such that $\frac{1}{e} \le \left(\frac{1-\epsilon}{\epsilon}\right)^{m-2\sum_{i=1}^m z_i} \le e$.

We denote the probability of the event $y \in \mathcal{Y}^1$ by $p_1 = \Pr\left(y \in \mathcal{Y}^1 \mid z_1,\dots,z_m\right)$, and similarly $p_0 = \Pr\left(y \in \mathcal{Y}^0 \mid z_1,\dots,z_m\right)$. Then we have $p_1 + p_0 = 1$ and $\frac{p_1}{p_0} = \frac{\Pr\left(z_1,\dots,z_m|y\in\mathcal{Y}^1\right)\Pr\left(y\in\mathcal{Y}^1\right)}{\Pr\left(z_1,\dots,z_m|y\in\mathcal{Y}^0\right)\Pr\left(y\in\mathcal{Y}^0\right)} = \frac{(1-\epsilon)^{\sum_{i=1}^m z_i}\epsilon^{m-\sum_{i=1}^m z_i}}{\epsilon^{\sum_{i=1}^m z_i}(1-\epsilon)^{m-\sum_{i=1}^m z_i}} = \left(\frac{\epsilon}{1-\epsilon}\right)^{m-2\sum_{i=1}^m z_i}$. Therefore, if $\frac{1}{e} \le \left(\frac{1-\epsilon}{\epsilon}\right)^{m-2\sum_{i=1}^m z_i} \le e$, then $p_1, p_0 \in \left[\frac{1}{4},\frac{3}{4}\right]$. Consider a test $Z_i \in \mathcal{Z}$, and assume the distribution on D_i is a Bernoulli(p) with $p \in \left[\frac{1}{4},\frac{1}{2}\right]$, then

$$\begin{aligned}\mathbb{I}(Z_i;Y) = \mathbb{I}(Z_i;D_i) &= h_2(p(1-\epsilon)+(1-p)\epsilon) - h_2(\epsilon) \\ &= \int_\epsilon^{p(1-\epsilon)+(1-p)\epsilon} h_2(x)'dx \\ &\ge \frac{h_2'(\epsilon)+h_2'(p(1-\epsilon)+(1-p)\epsilon)}{2}(p(1-\epsilon)+(1-p)\epsilon-\epsilon) \\ &= p(1-2\epsilon)\frac{\log\frac{1-\epsilon}{\epsilon}+\log\frac{1-p(1-2\epsilon)-\epsilon}{p(1-2\epsilon)+\epsilon}}{2} \\ &\ge \frac{1-2\epsilon}{4}\frac{\log\frac{1-\epsilon}{\epsilon}+\log\left(1+\frac{2(1-2\epsilon)}{1+2\epsilon}\right)}{2} \\ &\ge \frac{1-2\epsilon}{8}\left(\frac{1-2\epsilon}{1-\epsilon}+(1-2\epsilon)\right) \\ &\ge \frac{3}{8}(1-2\epsilon)^2.\end{aligned}$$

That is, when $p \in [1/4, 1/2]$, we have $\mathbb{I}(Z_i; Y) \geq \frac{3}{8}(1-2\epsilon)^2 > \frac{1}{4}S_{\min} > \frac{10+2\log\log n}{2\log n}$.

Also, given the assumptions of step 3, we note that the hypotheses in the set $\mathcal{Y}^1$ will always have equal probability. Therefore, the tests of other types have at most information twice as their gain in the very beginning where we have uniform distribution on $\mathcal{Y}$. Therefore, the greedy policy will certainly choose a test among $\mathcal{Z}$.

Finally, by combining Step 1-3, we finish the proof. ■