Near-optimal Bayesian Active Learning with Correlated and Noisy Tests

Yuxin Chen
ETH Zurich

S. Hamed Hassani
ETH Zurich

Andreas Krause
ETH Zurich

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 two problem instances, including a Bayesian experimental design task intended to distinguish among economic theories of how people make risky decisions, and an active preference learning task via pairwise comparisons.

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 (Dasgupta, 2004a; Settles, 2012), active feature evaluation[1] (Kaplan et al., 2005; Deshpande et al., 2014; Dasgupta, 2004a; Settles, 2012), Bayesian experimental design (Fedorov, 1972; Chaloner & Verdinelli, 1995), policy making (Heckerman et al., 1994; Runge et al., 2011), probabilistic planning and optimal control (Smallwood & Sondik, 1973), 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 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 (Chakaravarthy et al., 2007); 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 (Kosaraju et al., 1999). 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 (Chen et al., 2015a).

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

Appearing in Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

[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.

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. (2010) 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* (Golovin & Krause, 2011), 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 **E**quivalence **C**lass **E**dge **D**iscounting (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 independent, 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, and an active preference learning task via pairwise comparisons (§5). To facilitate better understanding, we provide the detailed proofs, illustrative examples and a third application on pool-based active learning in the supplemental material.

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

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).

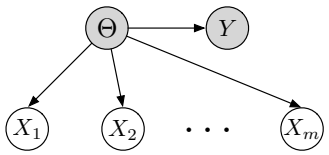


Figure 1: The basic model

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.

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

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., $\hat{y} \triangleq \arg\max_{y'\in\mathcal{Y}} \mathbb{P}\left[Y = y' \mid \psi_\pi\right]$. The error probability of predicting $\hat{y}$ is

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

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

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

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

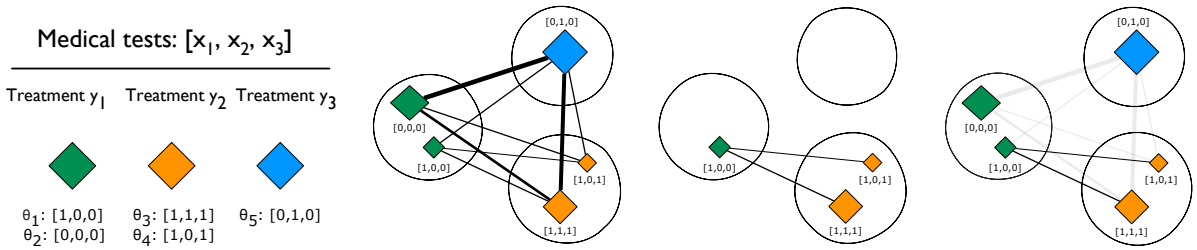
Computing the optimal policy for Problem (2.1) is intractable in general. When $\delta = 0$, this problem reduces to the equivalence class determination problem (Golovin et al., 2010; Bellala et al., 2010). Here, the target variables are referred to as *equivalence classes*, since each $y \in \mathcal{Y}$ corresponds to a subset of root-causes in $\mathrm{supp}(\Theta)$ that (equivalently) share the same "action".

Noise-free setting: the EC^2 algorithm. If tests are noise-free, i.e., $\forall e,\ \mathbb{P}\left[X_e \mid \Theta\right] \in \{0, 1\}$, this problem can be solved near-optimally by the *equivalence class edge cutting* (EC^2) algorithm (Golovin et al., 2010). 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 $w((\theta, \theta')) \triangleq \mathbb{P}\left[\theta\right] \cdot \mathbb{P}\left[\theta'\right]$, 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 $E(x_e) \triangleq \{\{\theta, \theta'\} \in E : \mathbb{P}\left[x_e \mid \theta\right] = 0 \ \vee\ \mathbb{P}\left[x_e \mid \theta'\right] = 0\}$ 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 $\mathbb{P}\left[\psi_2 \mid \psi_1\right] > 0$. 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 (Golovin & Krause, 2011).

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



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

Figure 2: (a) shows an illustrative example of the medical diagnosis problem. In (b), we initialize EC^2, by drawing edges between all pairs of root-causes (diamonds) 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 $X_1 = 1$. (d) ECED, instead, discounts the edge weights accordingly.

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. (2010) 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 the supplemental material (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.

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 with 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 purely noisy test X_1, i.e., $\forall\theta,\ \mathbb{P}[X_1 = 1 \mid \theta] = 0.5$, and (2) a noiseless test X_2 with $\mathbb{P}[X_2 = 1 \mid \theta_1] = 1$ and $\mathbb{P}[X_2 = 1 \mid \theta_2] = \mathbb{P}[X_2 = 1 \mid \theta_3] = 0$. One can easily verify that by running EC^2-Bayes, one actually prefers X_1 (with expected reduction in edge weight 0.18, as opposed to 0.112 for X_2).

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 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

[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.

EC^2-Bayes does), we choose to discount the root-causes by the *likelihood ratio*:

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

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. 2(d)). 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. $\mathbb{P}\left[X_e \mid \theta\right]$ is uniform, then $\lambda_{\theta,e} = 1$, 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:

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

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 $x_e \in \mathcal{O}$, $\mathbb{P}\left[X_e = x_e \mid \theta\right] = \mathbb{P}\left[X_e = x_e \mid \theta'\right]$. 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 $\delta_{\text{OFFSET}}(x_e \mid \psi_\pi) \triangleq \sum_{(\theta,\theta')\in E} \mathbb{P}\left[\theta, \psi_\pi\right] \mathbb{P}\left[\theta', \psi_\pi\right] \cdot (1 - \max_\theta \lambda^2_{\theta,x_e})$, as the *offset* value for observing outcome x_e. It is easy to check that if test e is non-informative, then it holds that $\delta_{\text{BS}}(x_e \mid \psi_\pi) - \delta_{\text{OFFSET}}(x_e \mid \psi_\pi) = 0$ for all $x_e \in \mathcal{O}$; otherwise $\delta_{\text{BS}}(x_e \mid \psi_\pi) - \delta_{\text{OFFSET}}(x_e \mid \psi_\pi) \geq 0$. This motivates us to use the following objective function:

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

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 depends on the number of root-causes. Let $\epsilon_{\theta,e} \triangleq 1 - \max_x \mathbb{P}\left[X_e = x \mid \theta\right]$ 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., $\forall \theta \in \text{supp}(\Theta)$, $\epsilon_{\theta,e} = \epsilon_e$), ECED is competitive with the optimal policy that achieves a lower error probability for Problem (2.1):

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

Input: $[\lambda_{\theta,x}]_{n\times m}$ (or Conditional Probabilities $\mathbb{P}\left[X \mid \Theta\right]$), Prior $\mathbb{P}\left[\Theta\right]$, Mapping $r : \text{supp}(\Theta) \to \mathcal{Y}$;
begin
  $\psi_\pi \leftarrow \emptyset$;
  foreach $(\theta, \theta') \in E$ **do**
    $w_{\theta,\theta'} \leftarrow \mathbb{P}\left[\theta\right] \mathbb{P}\left[\theta'\right]$;
  while $p^{\text{MAP}}_{\text{ERR}}(\psi_\pi) > \delta$ **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^*};
    $w_{\theta,\theta'} \leftarrow w_{\theta,\theta'} \cdot \mathbb{P}\left[x_{e^*} \mid \theta\right] \mathbb{P}\left[x_{e^*} \mid \theta'\right]$;
    $\psi_\pi \leftarrow \psi_\pi \cup \{(e^*, x_{e^*})\}$;
  Output: $y^* = \arg\max_y \mathbb{P}\left[y \mid \psi_\pi\right]$.

Theorem 1. *Fix $\delta \in (0,1)$. To achieve expected error probability less than δ, it suffices to run* ECED *for $O\left(\frac{k}{c_\varepsilon}\left(\log\frac{kn}{\delta}\log\frac{n}{\delta}\right)^2\right)$ steps where $n \triangleq |\text{supp}(\Theta)|$ denotes the number of root-causes, $c_\varepsilon \triangleq \min_{e\in\mathcal{V}}(1-2\epsilon_e)^2$ characterizes the severity of noise, and $k \triangleq \text{cost}\left(\mathsf{OPT}(\delta_{\mathsf{opt}})\right)$ is 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 $O\left(\left(\log(mn/\delta)\log(n/\delta)\right)^2/c_\varepsilon\right)$ 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 $p_{\text{ERR}} \leq \delta$. Finally, notice that by construction ECED never selects any non-informative test. Therefore, we can always remove purely noisy tests (i.e., $\{e : \forall\theta,\ \mathbb{P}\left[X_e = 1 \mid \theta\right] = \mathbb{P}\left[X_e = 0 \mid \theta\right] = 1/2\}$), 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 non-trivial 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 $f_{\text{AUX}} : 2^{\mathcal{V}\times\mathcal{O}} \to \mathbb{R}_{\geq 0}$, defined as

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

Here $\mathbb{H}_2(x) := -x\log x - (1-x)\log(1-x)$, 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 $n \triangleq |\operatorname{supp}(\Theta)|$ be the number of root-causes, and $p^{MAP}_{\text{ERR}}(\psi)$ be the error probability given partial realization ψ. Then*

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

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 $p^{\text{MAP}}_{\text{ERR}}$.

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 $\Delta_{\text{AUX}}(X_e \mid \psi) \triangleq \mathbb{E}_{x_e}[f_{\text{AUX}}(\psi\cup\{e,x_e\}) - f_{\text{AUX}}(\psi) \mid \psi]$ to be the expected gain of test e in f_{AUX}. Let $\Delta_{\text{EC}^2,\psi}(X_e)$ denote the gain of test e in the EC^2 objective, assuming that the edge weights are configured according to the *posterior distribution* $\mathbb{P}\left[\Theta \mid \psi\right]$. Similarly, let $\Delta_{\text{ECED},\psi}(X_e)$ denote the ECED gain, if the edge weights are configured according to $\mathbb{P}\left[\Theta \mid \psi\right]$. 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*

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

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 $\Delta_{\text{AUX}}(X_e \mid \psi_\pi)$, i.e., the gain in f_{AUX} under the *noisy* setting, to $\Delta_{\text{EC}^2,\psi}(X_e)$, 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:

$$\begin{aligned} f_{\text{EC}^2,\psi} &:= \sum_y \mathbb{P}\left[y \mid \psi\right]\left(1 - \mathbb{P}\left[y \mid \psi\right]\right) \\ &\overset{(a)}{\geq} 1 - \max_y \mathbb{P}\left[y \mid \psi\right] = p^{\text{MAP}}_{\text{ERR}}(\psi). \end{aligned} \tag{4.2}$$

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 $f_{\text{EC}^2,\psi}$ 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 the supplemental material.

5 Experimental Results

We now demonstrate the 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, and an active preference learning task via pairwise comparisons. For these two tasks, we run experiments on benchmark datasets which have been used in the existing literature. Due to space limitations, we defer a third case study on pool-based active learning to the Appendix.

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) (Howard, 1966). 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

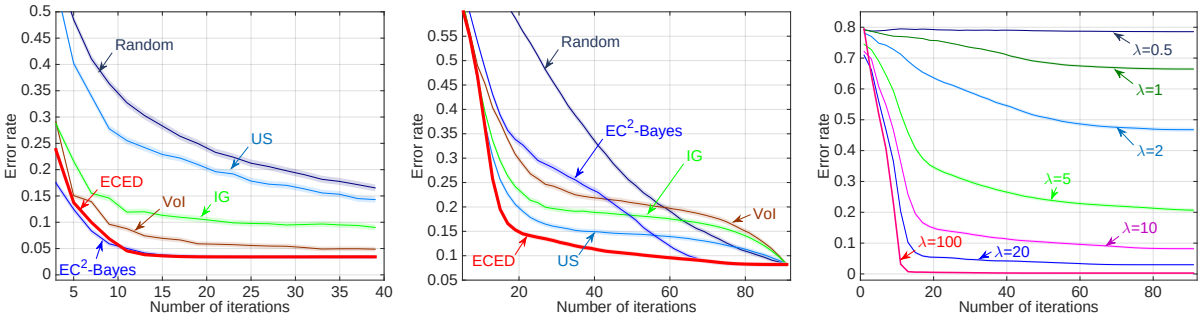
We first conduct experiments on a Bayesian experimental design task, which intends to distinguish among economic theories of how people make risky decisions. Several theories have been proposed in behavioral economics to explain how people make decisions under risk and uncertainty. We test ECED on six theories of subjective valuation of risky choices (Wakker, 2010; Tversky & Kahneman, 1992; Sharpe, 1964), 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 corresponds 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 Ray et al. (2012) 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] (Bradley & Terry, 1952), where the probability that root-cause θ favors L_1 is defined as $\mathbb{P}[X_e = 1 \mid \theta] = \frac{1}{1+\exp(-\lambda\cdot(v_1 - v_2))}$.

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. 3(a) 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 (Herlocker et al., 1999), which consists of a matrix of 1 to 5 ratings of 1682 movies from 943 users, and adopt the experimental setup proposed in Chen et al. (2015b). 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., $e \triangleq (m_a, m_b)$, where a and b are embeddings of movie m_a and m_b 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., $\mathbb{P}[X_e = 1 \mid \theta] = \frac{1}{1+\exp(-\lambda\cdot(d(m_a,\theta) - d(m_b,\theta)))}$. where $d(\cdot,\cdot)$ is the distance function, and λ controls the level of noise in the system.

[5] The BTL model has been widely used for pairwise data, e.g., Negahban et al. (2012); Shah et al. (2015), 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.



(a) *Risk Choice Theory* (b) *MovieLens* (c) *MovieLens* - Varying noise

Figure 3: Experimental results: ECED outperforms most baselines on both data sets.

Results. Fig. 3(b) 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. 3(c) 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, hence they are hardly informative in Y. It comes at the cost of performing more tests, and hence low convergence rate.

6 Related Work

Active learning in statistical learning theory. In most of the theoretical active learning literature (e.g., Dasgupta (2004b); Hanneke (2007, 2014); Balcan & Urner (2015)), active learning algorithms are mainly considered in terms of their *statistical complexity* (e.g., the reduction in labels required), disregarding their *computational complexity* (Balcan et al., 2006; Zhang & Chaudhuri, 2014). 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 (Dasgupta, 2004b), disagreement coefficient (Hanneke, 2007), etc); In comparison, in this paper 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 supp(Θ) 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 (Kääriäinen, 2006; Karp & Kleinberg, 2007; Nowak, 2009). 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 (Chen et al., 2015a) and strict noise assumptions (Golovin et al., 2010), 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. We have proved that ECED enjoys strong theoretical guarantees, by introducing an analysis framework that draws upon adaptive submodularity and information theory. 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.

8 Acknowledgments

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

References

Balcan, Maria-Florina and Urner, Ruth. Active learning–modern learning theory. *Encyclopedia of Algorithms*, 2015. 8

Balcan, Maria Florina, Beygelzimer, Alina, and Langford, John. Agnostic active learning. In *ICML*, pp. 65–72, 2006. 8

Bellala, G., Bhavnani, S., and Scott, C. Extensions of generalized binary search to group identification and exponential costs. In *NIPS*, 2010. 3

Bradley, Ralph Allan and Terry, Milton E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 7

Chakaravarthy, V. T., Pandit, V., Roy, S., Awasthi, P., and Mohania, M. Decision trees for entity identification: Approximation algorithms and hardness results. In *SIGMOD/PODS*, 2007. 1

Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. 1

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

Chen, Yuxin, Hassani, S. Hamed, Karbasi, Amin, and Krause, Andreas. Sequential information maximization: When is greedy near-optimal? In *COLT*, 2015a. 1, 8

Chen, Yuxin, Javdani, Shervin, Karbasi, Amin, Bagnell, James Andrew, Srinivasa, Siddhartha, and Krause, Andreas. Submodular surrogates for value of information. In *AAAI*, 2015b. 7

Dasgupta, S. Analysis of a greedy active learning strategy. In *NIPS*, 2004a. 1

Dasgupta, Sanjoy. Analysis of a greedy active learning strategy. In *NIPS*, 2004b. 8

Deshpande, Amol, Hellerstein, Lisa, and Kletenik, Devorah. 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*, pp. 1453–1467. SIAM, 2014. 1

Fedorov, Valerii Vadimovich. *Theory of optimal experiments*. Elsevier, 1972. 1

Golovin, Daniel and Krause, Andreas. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *JAIR*, 2011. 2, 3

Golovin, Daniel, Krause, Andreas, and Ray, Debajyoti. Near-optimal bayesian active learning with noisy observations. In *NIPS*, 2010. 2, 3, 4, 8

Hanneke, Steve. A bound on the label complexity of agnostic active learning. In *ICML*, 2007. 8

Hanneke, Steve. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7 (2-3):131–309, 2014. 8

Heckerman, David, Breese, John, and Rommelse, Koos. Troubleshooting under uncertainty. Technical report, Technical Report MSR-TR-94-07, Microsoft Research, 1994. 1

Herlocker, Jonathan L., Konstan, Joseph A., Borchers, Al, and Riedl, John. An algorithmic framework for performing collaborative filtering. In *SIGIR*, 1999. 7

Howard, R.A. Information value theory. *Systems Science and Cybernetics, IEEE Trans. on*, 2(1):22–26, 1966. 7

Kääriäinen, Matti. Active learning in the non-realizable case. In *Algorithmic Learning Theory*, pp. 63–77, 2006. 8

Kaplan, Haim, Kushilevitz, Eyal, and Mansour, Yishay. Learning with attribute costs. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pp. 356–365. ACM, 2005. 1

Karp, Richard M and Kleinberg, Robert. Noisy binary search and its applications. In *SODA*, 2007. 8

Kosaraju, S Rao, Przytycka, Teresa M, and Borgstrom, Ryan. On an optimal split tree problem. In *Algorithms and Data Structures*, pp. 157–168. Springer, 1999. 1

Negahban, Sahand, Oh, Sewoong, and Shah, Devavrat. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, pp. 2474–2482, 2012. 7

Nowak, Robert. Noisy generalized binary search. In *NIPS*, 2009. 8

Ray, Debajyoti, Golovin, Daniel, Krause, Andreas, and Camerer, Colin. Bayesian rapid optimal adaptive design (broad): Method and application distinguishing models of risky choice. *Tech. Report*, 2012. 7

Runge, M. C., Converse, S. J., and Lyons, J. E. Which uncertainty? using expert elicitation and expected value of information to design an adaptive program. *Biological Conservation*, 2011. 1

Settles, B. *Active Learning*. Morgan & Claypool, 2012. 1

Shah, Nihar B, Balakrishnan, Sivaraman, Bradley, Joseph, Parekh, Abhay, Ramchandran, Kannan, and Wainwright, Martin J. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *AISTATS*, 2015. 7

Sharpe, William F. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance*, 1964. 7

Smallwood, Richard D and Sondik, Edward J. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973. 1

Tversky, Amos and Kahneman, Daniel. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 1992. 7

Wakker, P.P. *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press, 2010. 7

Zhang, Chicheng and Chaudhuri, Kamalika. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems*, pp. 442–450, 2014. 8