Diameter-based Interactive Structure Discovery

Christopher Tosh
Columbia University

Daniel Hsu
Columbia University

Abstract

We introduce *interactive structure discovery*, a generic framework that encompasses many interactive learning settings, including active learning, top-k item identification, interactive drug discovery, and others. We adapt a recently developed active learning algorithm of Tosh and Dasgupta (2017) for interactive structure discovery, and show that the new algorithm can be made noise-tolerant and enjoys favorable query complexity bounds.

1 Introduction

Standard approaches to learning structures from data generally do not incorporate human interaction into the learning process. Typically, a data set is collected and labeled, if appropriate, and an algorithm is run to find the structure that best fits the data. *Interactive structure learning*, by contrast, adaptively solicits feedback from a human, or other information source, during the structure learning process. The hope is that by incorporating interaction into the learning process, we may be able to learn higher quality structures with less data or lower computational costs.

Recently, there has been interest in designing algorithms for interactive structure learning. Some works (Emamjomeh-Zadeh and Kempe, 2017; Tosh and Dasgupta, 2018) have attacked this problem in broad generality, designing algorithms that are capable of interactively learning generic classes of structures. Others have designed structure-specific interactive learning algorithms in a variety of settings, including flat and hierarchical clustering (Wagstaff and Cardie, 2000; Awasthi et al., 2014; Ashtiani et al., 2016; Vikram and Dasgupta, 2016), topic modeling (Hu et al., 2014; Lund et al., 2017), and matrix completion (Krishnamurthy and Singh, 2014). In all of these works, the ultimate goal is to find the structure that a user has in mind, and the algorithms are designed around this objective.

However, users of interactive learning algorithms are not always primarily interested in obtaining high-quality estimates of a particular structure. In some settings, especially those where actions are to be taken based on what has been learned, the goal is to glean information on some aspect of a structure. In information retrieval, for example, knowing the correct ranking of a set of items is often less important than getting the ordering of the first few elements correct (Mohajer et al., 2017; Shah and Wainwright, 2017).

In this work, we introduce *interactive structure discovery*, a general framework that encompasses both traditional interactive structure learning and other scenarios that have objectives which deviate from the structure estimation problem. We also demonstrate that there is a natural, general-purpose algorithm for this setting, and we give guarantees on its consistency and convergence rates, even in the presence of noise.

1.1 Paper organization

In Section 2, we introduce the problem of interactive structure discovery, and provide several examples illustrating the breadth of its potential applications. In Section 3, we introduce an algorithm, a generalization of the DBAL algorithm (Tosh and Dasgupta, 2017), for the interactive structure discovery problem. In Section 4, we show that this algorithm is consistent and enjoys fast rates of convergence under certain conditions. We also demonstrate nearly matching lower bounds. In Section 5, we provide concrete, worked examples of these theoretical guarantees. In particular, we illustrate the improvements that interactive structure discovery can offer over other schemes that focus purely on the standard structure estimation problem. We conclude in Section 6 with simulations demonstrating that the algorithms discussed here can be practically implemented and perform well on simulated data.

Proceedings of the 23$^{\mathrm{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

2 Interactive structure discovery

There are a variety of settings in which adaptively solicited interaction has been shown to decrease the statistical or computational resources required for a learning problem. In active learning, for example, algorithms that are able to adaptively query data points for their labels are able to find low-error classifiers with fewer labels than learning algorithms presented with random labels (Dasgupta, 2005; Balcan et al., 2010; Hanneke, 2011). In adaptive matrix completion, learners that adaptively query the entries of some unknown low-rank matrix are able to reconstruct the matrix with fewer revealed entries than can be done with randomly sampled entries (Krishnamurthy and Singh, 2014). In clustering, soliciting constraints from a user or oracle can improve the quality of the clustering (Vikram and Dasgupta, 2016) and circumvent computational hardness results (Ashtiani et al., 2016).

The examples above can be thought of as structure estimation problems – problems where the learner's objective is to estimate some ground-truth structure. However, there are also learning situations that can benefit from interaction but are not easily framed as structure estimation problems. In the top-k item identification problem, a learner queries the relative preferences of a user over a set of n items with the goal of finding the k most preferred items. While this problem can be solved by estimating a user's entire preference ordering, algorithms designed specifically for the top-k item identification problem can get away with fewer queries (Mohajer et al., 2017). Another interactive learning situation that is not so cleanly expressed as a structure estimation problem is the drug discovery problem (Barretina et al., 2012; Yang et al., 2012), which is much like the adaptive matrix completion problem except the goal is not to estimate the entire drug-cell interaction matrix, but rather it is to find a drug exhibiting certain properties.

In this section, we formalize the problem of *interactive structure discovery* which generalizes all of the above interactive learning settings into a single framework. Later, we will present a natural algorithm that operates within this general framework.

2.1 Structure decompositions

Denote by $\mathcal{G}$ the space of structures under consideration, these could be, for example, binary classifiers, or clusterings of some fixed data set, or low rank $n \times p$ matrices. Following Tosh and Dasgupta (2018), we view each structure in $\mathcal{G}$ as a function from a set of atomic questions $\mathcal{A}$ to a set of responses $\mathcal{Y}$. As the following examples illustrate, this view admits a wide spectrum of admissible structures.

- **Binary classifiers**. When $\mathcal{G}$ is a collection of classifiers, each atom $a \in \mathcal{A}$ corresponds to a data point and $\mathcal{Y} = \{0, 1\}$.
- **Clusterings**. If $\mathcal{G}$ is a set of clusterings of a collection of n items, then we may view $g \in \mathcal{G}$ as the function from $\mathcal{A} = \binom{[n]}{2}$ to $\mathcal{Y} = \{0, 1\}$, where $g((i,j))$ is 1 if i, j belong to the same cluster in g and 0 otherwise.
- **Binary hierarchical clusterings**. If $\mathcal{G}$ is a set of binary hierarchies over n items, then we may view $g \in \mathcal{G}$ as the function from $\mathcal{A} = \binom{[n]}{3}$ to $\mathcal{Y} = \{0, 1, 2\}$, where
$$g((i,j,k)) = \begin{cases} 0 & \text{if } i,j \text{ are clustered before } k \text{ in } g \\ 1 & \text{if } i,k \text{ are clustered before } j \text{ in } g \\ 2 & \text{if } j,k \text{ are clustered before } i \text{ in } g \end{cases}$$
- **Matrices**. If $\mathcal{G}$ is a set of $n \times p$ matrices, then $\mathcal{A} = [n] \times [p]$ and $\mathcal{Y} = \mathbb{R}$, and $g((i,j))$ is the (i,j)-th entry of the matrix corresponding to g.

We will assume that there is some distribution $\mathcal{D}$ over $\mathcal{A}$. In the case of classifiers, $\mathcal{D}$ is the data distribution. For clusterings over a fixed collection of items or matrices of a fixed size, a reasonable choice for $\mathcal{D}$ would be the uniform distribution over $\mathcal{A}$.

2.2 Structure distances

We are interested in settings where the goal may not be to recover a particular structure but perhaps only to recover some aspect of that structure. We capture this objective in the form of a *structure distance* $d : \mathcal{G} \times \mathcal{G} \to \mathbb{R}_{\geq 0}$, which we assume to be positive, symmetric, and satisfy $d(g,g) = 0$ for all $g \in \mathcal{G}$. In particular, we do not require this structure distance to satisfy the triangle inequality. If $g^* \in \mathcal{G}$ is a ground-truth structure, then our objective is to find a structure $g \in \mathcal{G}$ such that $d(g, g^*)$ is small. We illustrate the flexibility of this approach with some examples.

- **Low-error classifiers**. If our objective is to find a classifier with low error, then we make take our distance to be
$$d(g, g') \;=\; \Pr_{a \sim \mathcal{D}}(g(a) \neq g'(a)).$$
A classifier g satisfying $d(g, g^*) < \epsilon$ will have error less than ϵ. For this reason, this is the standard classification distance used to learn low-error classifiers in active learning. More generally, this is a reasonable notion of distance if our goal is to learn a high quality structure (Tosh and Dasgupta, 2018).

- **Fair classifiers**. A recently proposed notion of fairness, called equal opportunity (Hardt et al., 2016), attempts to balance the number of true positives between individuals with a certain protected attribute and those without the protected attribute. If our goal is to find a classifier that approximately satisfies this notion of fairness while simultaneously achieving low error, then we may take $d(g, g')$ to be

  $$\begin{aligned}&\max\{\Pr_{a\sim\mathcal{D}}(g(a) \neq g'(a)),\\ &\quad\lambda|\mathbb{E}_{a\sim\mathcal{D}_0}[g(a)|g'(a)=1] - \mathbb{E}_{a\sim\mathcal{D}_1}[g(a)|g'(a)=1]|,\\ &\quad\lambda|\mathbb{E}_{a\sim\mathcal{D}_0}[g'(a)|g(a)=1] - \mathbb{E}_{a\sim\mathcal{D}_1}[g'(a)|g(a)=1]|\}\end{aligned}$$

  where D_p denotes the distribution of a point conditioned on it having protected attribute value p and $\lambda > 0$ is some weight of the relative importance of fairness. If we find a g satisfying $d(g, g^*) < \epsilon$, then we know that the error of g is at most ϵ and we violate equal opportunity by at most ϵ/λ.

- **Cluster identification**. In certain clustering situations, there is some particular item of interest i^*, and our goal is to find the cluster to which i^* belongs. In this case, we may take $d(g, g')$ to be

  $$\max\left\{\frac{|C(g,i^*) \setminus C(g',i^*)|}{|C(g,i^*)|}, \frac{|C(g',i^*) \setminus C(g,i^*)|}{|C(g',i^*)|}\right\}$$

  where $C(g, i) = \{j \in [n] : g((i, j)) = 1\}$ is the set of items in the same cluster as i under g. If we find a g satisfying $d(g, g^*) < \epsilon$, then we know that $C(g, i^*)$ is missing at most an ϵ fraction of the elements of $C(g^*, i^*)$ and at most an ϵ fraction of $C(g, i^*)$ is not included in $C(g^*, i^*)$.

- **Column selection**. If our goal is to find the best column of an $n \times p$ matrix as measured by some score function $s : \mathbb{R}^n \to \mathbb{R}$, then we may define our distance as

  $$\begin{aligned}d(g, g') \;=\; \max\{&s(g(\cdot, j_g)) - s(g(\cdot, j_{g'})),\\ &s(g'(\cdot, j_{g'})) - s(g'(\cdot, j_g))\}\end{aligned}$$

  where $g(\cdot, j)$ denotes the jth column of g and $j_g = \arg\max_j s(g(\cdot, j))$. If we find a g satisfying $d(g, g^*) < \epsilon$ and select column j_g, then the true score of j_g is at most ϵ worse then the true score of the best column.

As the preceding examples show, the structure distance is a flexible way to encode objectives into the structure discovery problem. Throughout the remainder of the paper, we will assume that we have such a distance $d(\cdot, \cdot)$, that our objective is to find a $g \in \mathcal{G}$ satisfying $d(g, g^*) < \epsilon$ for some $\epsilon > 0$, and that we can efficiently compute $d(g, g')$ for any two structures $g, g' \in \mathcal{G}$. We will also assume that $d(g, g') \leq 1$, which can be achieved with an appropriate normalization.

3 Diameter-based structure discovery

Given a set of structures $\mathcal{G}$ and a suitable distance, how do we find a structure with low distance to the ground truth? One approach, which Tosh and Dasgupta (2017) proposed for the realizable binary classification setting, is to try to find a distribution over $\mathcal{G}$ such that structures are close to g^* on average. We take up their approach again here in our more general and potentially noisy setting.

Let π be some probability measure over $\mathcal{G}$. Define the *average diameter* of π as

$$\text{avg-diam}(\pi) \;=\; \mathbb{E}_{g,g'\sim\pi}[d(g, g')].$$

The following result, due to Tosh and Dasgupta (2017), shows that if one can find a distribution π with low average diameter that puts sufficient mass on a target structure g^*, then one can readily find a structure with small distance to g^* by random sampling.

Lemma 1. *If $g^* \in \mathcal{G}$ and π is a distribution over $\mathcal{G}$, then $\mathbb{E}_{g\sim\pi}[d(g, g^*)] \leq \text{avg-diam}(\pi)/\pi(g^*)$.*

Although Lemma 1 was originally stated for the case where $d(\cdot, \cdot)$ is the disagreement probability of two classifiers, it still holds in our setting.

Lemma 1 reduces the problem of finding a structure close to g^* to that of finding a distribution π with low average diameter, provided we can sample from it. Thus, we are interested in queries whose answers will help us find distributions with low average diameter. This motivates the concept of average splitting.

For any subset $V \subset \mathcal{G}$, let $\pi|_V$ denote the conditional distribution of π restricted to V. For a given atom $a \in \mathcal{A}$ and a possible response $y \in \mathcal{Y}$, let $\mathcal{G}_a^y = \{g \in \mathcal{G} : g(a) = y\}$ denote the set of structures consistent with y on atom a. For any $a \in \mathcal{A}$, we say that a *ρ-average splits* π if

$$\max_{y\in\mathcal{Y}} \pi(\mathcal{G}_a^y)^2 \,\text{avg-diam}(\pi|_{\mathcal{G}_a^y}) \leq (1-\rho)\,\text{avg-diam}(\pi) \quad (1)$$

We say that π is (ρ, τ)-*average splittable* if the probability that a random a drawn from $\mathcal{D}$ ρ-average splits π is at least τ; and we say that $\mathcal{G}$ has *average splitting index* (ρ, ϵ, τ) if any distribution π over $\mathcal{G}$ satisfying $\text{avg-diam}(\pi) > \epsilon$ is (ρ, τ)-average splittable.

Given an efficient sampler for π, we can estimate all of the relevant quantities in equation (1) via Monte Carlo approximations: if g, g' are drawn i.i.d. from π then for any $a \in \mathcal{A}$ and $y \in \mathcal{Y}$,

$$\mathbb{E}[d(g, g')\mathbb{1}[g(a) = y = g'(a)]] = \pi(\mathcal{G}_a^y)^2\text{avg-diam}(\pi|_{\mathcal{G}_a^y}).$$

3.1 Finding a good query

Suppose that we want to choose from a set of atoms the one that provides the largest average split, say ρ, of π. How do we go about doing this? In the case where $\mathcal{G}$ is a binary hypothesis class and avg-diam(π) has a known lower bound ϵ, Tosh and Dasgupta (2017) gave an algorithm that can find a query that $O(\rho)$-average splits π while sampling $\tilde{O}(1/(\epsilon\rho^2) + 1/\text{avg-diam}(\pi)^2)$[1] structures from π.

In Algorithm 2, we present an algorithm based on inverse sampling (Haldane, 1945) that enjoys the same guarantees in a more general setting while sampling fewer structures.

Lemma 2. *Pick $\alpha, \delta > 0$. If* SELECT *is run with atoms $a_1, \ldots, a_m$, one of which ρ-average splits π, then with probability $1-\delta$,* SELECT *returns a data point that $(1-\alpha)\rho$-average splits π while sampling no more than*

$$\frac{12}{\alpha^2(1-\alpha)\rho\,\text{avg-diam}(\pi)} \log\frac{m+|\mathcal{Y}|}{\delta}$$

pairs of structures in total.

With high probability, the running time of SELECT is $O\left(\frac{T_{\text{sample}}+m|\mathcal{Y}|}{\alpha^2(1-\alpha)\rho\,\text{avg-diam}(\pi)}\log\frac{m+|\mathcal{Y}|}{\delta}\right)$, where T_{sample} is the time needed to sample a structure.

For space reasons, all of the proofs in the paper are deferred to the appendix, but we sketch the main intuition of SELECT here. Say that $g, g' \sim \pi$. The key observation is that each atom a_i has an associated average split ρ_i such that for any response y,

$$\mathbb{E}[d(g,g')(1-\mathbb{1}[g(a_i)=y=g'(a_i)])] \geq \rho_i\,\text{avg-diam}(\pi)$$

and moreover there exists some response y^* such that

$$\mathbb{E}[d(g,g')(1-\mathbb{1}[g(a_i)=y^*=g'(a_i)])] = \rho_i\,\text{avg-diam}(\pi).$$

Suppose that we choose N and draw $g_j, g'_j \sim \pi$ sequentially until a round K_i in which all $y \in \mathcal{Y}$ satisfy

$$S_{K_i}^{a_i,y} = \sum_{j=1}^{K_i} d(g_j,g'_j)(1-\mathbb{1}[g_j(a_i)=y=g'_j(a_i)]) \geq N.$$

Then one can show that K_i is tightly concentrated around $\frac{N}{\rho_i\,\text{avg-diam}(\pi)}$ (Haldane, 1945).

Thus, the first atom a_i to satisfy that $S_K^{a_i,y} \geq N$ is likely to satisfy that $\rho_i \geq (1-\alpha)\max_j \rho_j$ for some constant α and the number of rounds needed for this to happen will satisfy $K \approx \frac{N}{\rho_i\,\text{avg-diam}(\pi)}$.

3.2 Noise-tolerant DBAL

The approach of Tosh and Dasgupta (2017) was to maintain a distribution π_t over all structures that are consistent with the feedback observed so far. In our setting, this corresponds to the posterior update rule

$$\pi_t(g) \;\propto\; \pi_{t-1}(g)\mathbb{1}[g(a_t)=y_t] \qquad (2)$$

after querying a_t and receiving response y_t. Their algorithm, termed DBAL for Diameter-based Active Learning, was shown to have favorable query complexity dependence on the average splitting index in the noiseless and realizable binary classification setting.

In this work, we want to be able to handle settings where our responses are noisy or inconsistent with a ground-truth structure. Following Nowak (2011), we consider a 'softer' posterior update:

$$\pi_t(g) \;\propto\; \pi_{t-1}(g)\exp(-\beta\mathbb{1}[g(a_t)\neq y_t]) \qquad (3)$$

where $\beta > 0$ is some parameter corresponding roughly to our confidence in the accuracy of the responses. Note that by taking $\beta \to \infty$, we recover the update in equation (2). We call this algorithm NDBAL for Noise-tolerant Diameter-based Active Learning. The full algorithm for NDBAL is displayed in Algorithm 1.

The update in equation (3) has been shown to enjoy favorable guarantees for active learning strategies that attempt to shrink π-mass (Nowak, 2011; Tosh and Dasgupta, 2018). We will show that it also works well for NDBAL.

4 Theoretical guarantees

In this section, we establish the statistical consistency of NDBAL and study its rate of convergence. To do so, we need to formalize our problem set up. Note that at each time t, the random outcomes consist of the atom a_t that we query, as well as the response y_t to a_t. Let $\mathcal{F}_t$ denote the sigma-field of all outcomes up to and including time t.

4.1 Consistency

We first show that NDBAL is consistent, i.e. $\mathbb{E}_{g\sim\pi_t}[d(g,g^*)] \to 0$ as $t \to \infty$ almost surely (a.s.), where $g^* \in \mathcal{G}$ is a ground truth structure. To do so, we need to make a few assumptions on our problem set up. Our first assumption is that $\mathcal{G}$ is finite. This will be relaxed when we study faster rates of convergence.

Our next assumption is that any two structures with positive distance can be distinguished by a random atom with positive probability.

Assumption 1. *For any $g, g' \in \mathcal{G}$ such that $d(g,g') > 0$, we have $\Pr_{a\sim\mathcal{D}}(g(a) \neq g'(a)) > 0$.*

[1] The $\tilde{O}(\cdot)$ suppresses logarithmic factors in $1/\delta$ and the number of candidate atoms.

Algorithm 1 NDBAL

```
Input: Distribution $\pi$, $\beta > 0$, $\alpha, \delta \in (0,1)$
Initialize $\pi_o = \pi$
for $t = 1, 2, \ldots$ do
  Draw $m$ atoms $\mathbf{a} = (a_1, \ldots, a_m)$
  Query $a_t = \text{SELECT}(\pi_{t-1}, \mathbf{a}, \alpha, \delta)$ and receive $y_t$
  $\pi_t(g) \propto \pi_{t-1}(g) \exp\left(-\beta \mathbb{1}[g(a_t) \neq y_t]\right)$
end for
return Posterior $\pi_t$
```

Algorithm 2 SELECT

```
Input: Distribution $\pi$, atoms $a_1, \ldots, a_m$
Set $N = \frac{6(2+\alpha)}{\alpha^2} \ln \frac{m+|\mathcal{Y}|}{\delta}$, $K = 0$, $S_0^{a_i,y} = 0$
for $K = 1, 2, \ldots$ do
  Draw $g, g' \sim \pi$ and compute for all $a_i, y$:
    $S_K^{a_i,y} = S_{K-1}^{a_i,y} + d(g,g')(1 - \mathbb{1}[g(a_i) = y = g'(a_i)])$
  If $\exists a_i$ s.t. $S_K^{x_i,y} \geq N$ for all $y \in \mathcal{Y}$, halt and return $a_i$.
end for
```

Note that Assumption 1 is necessary for identifiability: when Assumption 1 does not hold, there exist structures g, g' with $d(g,g') > 0$ that cannot be distinguished with atomic questions.

We will also need to make an assumption on the typical responses provided by a user. Let $\eta(y \,|\, a)$ denote the conditional probability of response y to atomic question a. We will require that the most likely response to an atomic query is the true response.

Assumption 2. *There exist $g^* \in \mathcal{G}$ and $\lambda > 0$ s.t.*

$$\eta(g^*(a) \,|\, a) \geq \eta(y \,|\, a) + \lambda$$

for any $a \in \mathcal{A}$ and $y \neq g^*(a)$.

In the setting where $\mathcal{G}$ is a collection of binary classifiers, Assumption 2 is equivalent to Massart's bounded noise condition (Awasthi et al., 2015). This noise condition has been previously studied in the active learning literature under the related notion of the splitting index (Balcan and Hanneke, 2012, Appendix C), albeit with a different active learning algorithm.

Our analysis will focus on the behavior of the potential function avg-diam$(\pi_t)/\pi_t(g^*)$. By Lemma 1, whenever this potential function goes to 0, $\mathbb{E}_{g\sim\pi_t}[d(g,g^*)]$ also must go to 0. The following lemma demonstrates that under Assumption 2, a related potential function is guaranteed to decrease in expectation.

Lemma 3. *Pick $k \geq 2$. Suppose Assumption 2 holds and $\beta \leq \lambda/(2+2k^2)$. If we query an atom a_t that ρ-average splits π_{t-1}, then in expectation over the randomness of the response y_t, we have*

$$\mathbb{E}\left[\frac{\text{avg-diam}(\pi_t)}{\pi_t(g^*)^k}\,\middle|\,\mathcal{F}_{t-1}, a_t\right] = (1-\Delta)\,\frac{\text{avg-diam}(\pi_{t-1})}{\pi_{t-1}(g^*)^k}$$

where $\Delta \geq \rho\lambda\beta/2$.

Thus, at each at each round, avg-diam$(\pi_t)/\pi_t(g^*)^k$ decreases in expectation by a multiplicative factor of $1-\Delta$, for an appropriate choice of β. However, Lemma 3 does not tell us how avg-diam(π_t) and $\pi_t(g^*)$ behave individually. The following lemma shows that $1/\pi_t(g^*)^k$ is a supermartingale.

Lemma 4. *Pick $k \geq 1$. Suppose Assumption 2 holds and $\beta \leq \lambda/k$. Then for any query a_t, we have $\mathbb{E}\left[1/\pi_t(g^*)^k \,|\, \mathcal{F}_{t-1}, a_t\right] \leq 1/\pi_{t-1}(g^*)^k$.*

Lemma 3 also tells us how much avg-diam$(\pi_t)/\pi_t(g^*)^k$ decreases in expectation given that we query a point that ρ-average splits the current posterior. In order to demonstrate consistency, we need ρ to be lower bounded on average. The following lemma gives such a lower bound for points chosen by NDBAL.

Lemma 5. *If Assumption 1 holds and NDBAL is run with constants $\alpha, \delta \in (0,1)$, then there is a constant $c > 0$, depending on $\alpha, \delta, d(\cdot,\cdot), \mathcal{G}$ and $\mathcal{D}$, such that for every round t, NDBAL queries a point that ρ_t-average split π_t satisfying $\mathbb{E}[\rho_t \,|\, \mathcal{F}_{t-1}] \geq \frac{c}{1-\log(\text{avg-diam}(\pi_t))}$.*

In the appendix, we show how the above results imply consistency for NDBAL.

Theorem 6. *If Assumptions 1 and 2 hold, $\beta \leq \lambda/10$, and $\pi_o(g^*) > 0$, then $\mathbb{E}_{g\sim\pi_t}[d(g,g^*)] \to 0$ a.s.*

4.2 Convergence rates

We now turn to the setting where there is some fixed error threshold $\epsilon > 0$, and our goal is to find a distribution π_t satisfying $\mathbb{E}_{g\sim\pi_t}[d(g,g^*)] \leq \epsilon$. The following theorem gives a bound on the resources that NDBAL uses to find such a distribution.

Theorem 7. *Let $\epsilon, \delta > 0$ and $\epsilon_o = \epsilon\delta\pi(g^*)/4$. If Assumption 2 holds, $\mathcal{G}$ has average splitting index (ρ, ϵ_o, τ) and NDBAL is run with $\beta \leq \lambda/10$ and $\alpha = 1/2$, then with probability $1-\delta$, NDBAL encounters a distribution π_t satisfying $\mathbb{E}_{g\sim\pi_t}[d(g,g^*)] \leq \epsilon$ while the resources used satisfy:*

(a) $T \leq \frac{2}{\rho\lambda\beta(1-\beta)} \max\left(\ln\frac{1}{\epsilon\pi(g^*)^2}, \frac{2e^{2\beta}}{\rho\lambda\beta(1-\beta)} \ln\frac{1}{\delta}\right)$ rounds, with one query per round,

(b) $m_t \leq \frac{1}{\tau} \log\frac{4t(t+1)}{\delta}$ atoms drawn per round, and

(c) $n_t \leq O\left(\frac{1}{\rho\epsilon_o} \log\frac{(m_t+|\mathcal{Y}|)t(t+1)}{\delta}\right)$ structures sampled per round.

While Theorem 7 does provide rates of convergence, it has several issues.

(i) The number of structures sampled in each round is polynomial in $1/\pi(g^*)$, which can be large.

(ii) Theorem 7 only guarantees that some posterior we encounter will satisfy avg-diam$(\pi_t)/\pi_t(g^*)^2 < \epsilon$; in particular, it does not tell us how to detect *which* posterior satisfies this property.

(iii) The average splitting index (ρ, ϵ_o, τ) depends on $\pi(g^*)$. In settings where the average splitting index has been bounded (Dasgupta, 2005; Tosh and Dasgupta, 2017), ρ and τ depend on ϵ_o, implying that the query complexity and the number of atoms drawn per round grow as $\pi(g^*)$ shrinks.

Without any further assumptions, issues (i) and (iii) are unavoidable even in the noiseless setting. To see why, consider a setting in which our prior only puts mass on two structures g and g^* where $d(g, g^*) \approx 1$. If structures are only accessed via a sampling oracle, detecting that there are two structures with positive probability mass requires $\Omega(1/\pi(g^*))$ samples. Moreover, in this scenario we have $\mathbb{E}_{g' \sim \pi}[d(g', g^*)] > \epsilon$ whenever avg-diam$(\pi)/\pi(g^*) > \epsilon/2$. Thus, with no further assumptions, we need to incur computational and data complexity costs that depend on $\pi(g^*)$.

4.3 Faster convergence rates

As discussed above, when g^* is completely independent of our prior π, NDBAL incurs high computational and data complexity costs. We show that this is avoided under the following Bayesian assumption on g^*.

Assumption 3. *There exists a $\lambda \geq 1$ and distribution ν over $\mathcal{G}$ such that the true structure g^* is drawn from ν and $1/\lambda \leq \nu(g)/\pi(g) \leq \lambda$ for every $g \in \mathcal{G}$.*

Assumption 3 is a slight relaxation of the traditional Bayesian assumption. Here we do not require g^* to be drawn from π itself, but rather only that it is drawn from some distribution that is close to π.

For ease of presentation, we also assume that we are in the completely noiseless setting. In the appendix, we show that we there is a certain amount of noise that we can tolerate and still get very fast rates of convergence. Formally, we make the following assumption.

Assumption 4. *$\exists g^* \in \mathcal{G}$ such that $\eta(g^*(a) \,|\, a) = 1$.*

With Assumption 4, we will run NDBAL with $\beta = \infty$ and get the posterior update in equation (2).

Together, Assumptions 3 and 4 immediately add more structure to our setting. In particular, if we have query/response pairs $(a_1, y_1), \ldots, (a_t, y_t)$, then the true posterior takes the form

$$\nu_t(g) \;\propto\; \nu(g)\mathbb{1}[g(a_i) = y_i \text{ for } i = 1, \ldots, t].$$

Without access to ν, there is no way to compute ν_t directly; however, we may still hope that a random draw from our distribution π_t is close to a random draw from ν_t, i.e. that the quantity

$$D(\pi_t, \nu_t) \;=\; \mathbb{E}_{g \sim \pi_t, g^* \sim \nu_t}[d(g, g^*)]$$

is small. Thus, our new objective is to find a distribution π_t satisfying $D(\pi_t, \nu_t) \leq \epsilon$. Given this new objective, we relax the requirement that $\mathcal{G}$ is finite. Instead, we assume that $\mathcal{G}$ has bounded *graph dimension* (Natarajan, 1989), a multiclass generalization of the VC dimension.

Definition 8. *Let $S = \{a_1, \ldots, a_m\}$ be a set of atomic questions. We say $\mathcal{G}$* shatters *S if there exists $f : S \to \mathcal{Y}$ such that for all $T \subset S$, there exists $g_T \in \mathcal{G}$ such that $g_T(x) = f(x)$ when $x \in T$ and $g_T(x) \neq f(x)$ when $x \in S \setminus T$. The* graph dimension *of $\mathcal{G}$ is the size of the largest S such that $\mathcal{G}$ shatters S.*

Finally, we need to decide when to stop making queries. As discussed in the previous section, one of the shortcomings of Theorem 7 is that it gives no guidance on when we have found a good distribution π_t. To address this, we use the stopping rule suggested by Tosh and Dasgupta (2017): estimate avg-diam(π_t) by sampling $\tilde{O}(\lambda^2/\epsilon)$ pairs of structures at the beginning of each round and stop if this estimate is below $3\epsilon/(4\lambda^2)$. Given this modification, we can improve the guarantees of NDBAL.

Theorem 9. *Suppose $\mathcal{G}$ has average splitting index $(\rho, \epsilon/(2\lambda^2), \tau)$ and graph dimension d_G. If Assumptions 3 and 4 hold, then with probability $1 - \delta$, modified* NDBAL *terminates with a distribution π_t satisfying $D(\pi_t, \nu_t) \leq \epsilon$ while using the following resources:*

(a) $T \leq O\left(\frac{d_G}{\rho}\left(\log\frac{|\mathcal{Y}|\lambda}{\epsilon\tau\delta} + \log^2\frac{d_G}{\rho}\right)\right)$ rounds with one query per round,

(b) $m_t \leq O\left(\frac{1}{\tau}\log\frac{t}{\delta}\right)$ atoms drawn per round, and

(c) $n_t \;\leq\; O\left(\left(\frac{\lambda^2}{\epsilon\rho}\right)\log\frac{(m_t + |\mathcal{Y}|)t}{\delta}\right)$ structures sampled per round.

In the appendix, we also consider the noisy setting.

4.4 Lower bounds

The results above demonstrate that the average splitting index provides upper bounds on the resource complexity of NDBAL in this generic interactive structure discovery setting. The following theorem shows that, in fact, some dependence on the average splitting index is inevitable for *any* learner in this setting.

Theorem 10. *Fix $\mathcal{G}$, $\mathcal{D}$ and $d(\cdot, \cdot)$. If $\mathcal{G}$ does not have average splitting index $(\frac{\rho}{4\lceil\log 1/\epsilon\rceil}, 2\epsilon, \tau)$ for some $\rho, \epsilon \in (0, 1)$ and $\tau \in (0, 1/2)$, then any interactive learning strategy which with probability $> 3/4$ over the random sampling from $\mathcal{D}$ finds a structure $g \in \mathcal{G}$ within distance $\epsilon/2$ of any target in $\mathcal{G}$ must draw at least $1/\tau$ atoms from $\mathcal{D}$ or must make at least $1/\rho$ queries.*

The proof of Theorem 10 is similar to the one by Dasgupta (2005) for lower bounding active learning, but adjusted to our more general setting.

5 Illustrative examples

In this section, we look at two specific structure learning settings. The first setting is the problem of learning a ranking over objects with features, where we provide bounds on the average splitting index. Combined with the results from Section 4, this gives us bounds on the performance of NDBAL.

The second setting is the problem of clustering the real line into k intervals. Here we demonstrate that the choice of structure distance can greatly influence the number of queries needed. In particular, when the structure distance only concerns a constant number of clusters, the label complexity of interactive structure discovery can be far smaller than when a more generic distance depending on the whole structure is used.

5.1 Feature-based rankings

In feature-based ranking, we have distribution μ over objects, each with corresponding feature vector $x \in \mathbb{R}^d$. A ranking corresponds to a weight vector $w \in \mathcal{G} = S^{d-1}$ (the unit sphere), where w ranks x over y if and only if $\langle w, x \rangle > \langle w, y \rangle$, in which case we write $w(x, y) = 1$ and 0 otherwise.

A natural ranking distance here is the following generalization of the Kendall tau distance:

$$d_r(w, w') = \Pr_{x,y\sim\mu}(w(x, y) \neq w'(x, y)).$$

The following theorem bounds the average splitting index when μ is spherically symmetric.

Theorem 11. *Suppose μ is spherically symmetric. Under distance $d_r(\cdot, \cdot)$, $\mathcal{G}$ has average splitting index $(\frac{1}{16\lceil\log(2/\epsilon)\rceil}, \epsilon, c\epsilon)$ for some absolute constant $c > 0$.*

Combining Theorem 11 with Theorems 7 and 9, the label complexity of NDBAL in this setting grows polylogarithmically in $1/\epsilon$.

5.2 Clustering on the line

Consider the problem of clustering the real line into k intervals where there is some interval $\mathcal{I}$ that we know should be clustered together under the ground truth clustering, and our goal is to identify the other points on the line that should be clustered with $\mathcal{I}$.

Say there is some measure μ over the real line, and let $\mathcal{G}_{k,\mathcal{I}}$ denote the set of clusterings of the real line into $\leq k$ intervals such that $\mathcal{I}$ is contained completely in one of these intervals. Note that a clustering $g \in \mathcal{G}_{k,\mathcal{I}}$ can be described by $k - 1$ reals $a_1 \leq a_2 \leq \cdots \leq a_{k-1}$.

The atomic questions consist of pairs of points (x, y), where $g(x, y) = 1$ if they belong to the same cluster and 0 otherwise. A natural distribution $\mathcal{D}$ over atomic questions is the product distribution $\mu \otimes \mu$, and a natural clustering distance is given by

$$d_c(g, g') = \Pr_{x,y\sim\mu}(g(x, y) \neq g'(x, y)).$$

However, if our goal is to identify the cluster that $\mathcal{I}$ belongs to, then a more intuitive clustering distance to use is given by

$$d_{\mathcal{I}}(g, g') = \Pr_{x\sim\mu}(g(x, \mathcal{I}) \neq g'(x, \mathcal{I}))$$

where $g(x, \mathcal{I}) = g(x, z)$ for all $z \in \mathcal{I}$.

Given these two notions of clustering distance, as well as our underlying goal of identifying the cluster that $\mathcal{I}$ belongs to, we ask whether there is a query complexity improvement in using an interactive structure discover algorithm such as NDBAL with distance $d_{\mathcal{I}}(\cdot, \cdot)$ as opposed to just learning with the standard clustering distance $d_c(\cdot, \cdot)$. Informally, we show the following.

Theorem 12 (Informal statement)**.** *There are settings in which learning under distance $d_c(\cdot, \cdot)$ with any interactive learning algorithm requires exponentially more queries than learning under $d_{\mathcal{I}}(\cdot, \cdot)$ with* NDBAL.

To prove Theorem 12, we derive the following bound on the average splitting index under distance $d_{\mathcal{I}}(\cdot, \cdot)$.

Lemma 13. *Let $\mu(\mathcal{I}) = \alpha$. Under distance $d_{\mathcal{I}}(\cdot, \cdot)$, $\mathcal{G}_{k,\mathcal{I}}$ has average splitting index $(\frac{1}{16\lceil\log(2/\epsilon)\rceil}, \epsilon, \frac{\epsilon\alpha}{2})$.*

6 Simulations

We now turn to experimentally evaluating NDBAL in two settings: linear classifiers and logit choice models. Before doing so, we discuss a modification to NDBAL that allows it to be run in practice.

General-loss NDBAL While the posterior update in Equation (3) enjoys nice theoretical properties, it results in a posterior distribution that may be intractable to sample from. Thus, we consider a more general update:

$$\pi_t(g) \;\propto\; \pi_{t-1}(g) \exp(-\beta\ell(g(a_t), y_t)) \qquad (4)$$

where $\ell(\cdot, \cdot)$ is some loss function. When the prior distribution π is log-concave, the loss function is convex, and $\mathcal{G}$ is convex, this results in a posterior distribution that is log-concave, and thus efficiently samplable (Lovasz and Vempala, 2007). Moreover, this update was shown to enjoy nice consistency properties for interactive learning strategies that query high variance atoms (Tosh and Dasgupta, 2018).

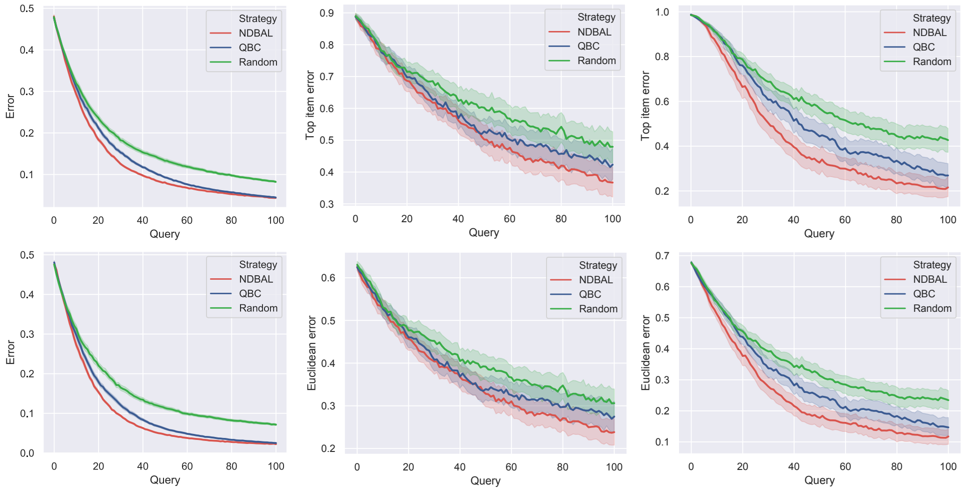

Figure 1: *Left*: Logistic noise simulations with $d = 10$. [*Top to bottom*: $\sigma = 5, 10$]. *Center and right*: Logit choice model experiments with $\sigma = 1, 5$. [*Top*: Top-item error. *Bottom*: distance to best item error.]

To formalize this setting, let $\mathcal{Y}$ denote the space of answers to atomic questions $\mathcal{A}$, and let $\mathcal{Z} \subset \mathbb{R}^d$ denote some prediction space for structures in $\mathcal{G}$. We view each structure in $\mathcal{G}$ as a function from $\mathcal{A}$ to $\mathcal{Z}$, and we suffer loss $\ell(z, y)$ for predicting z given answer y.

Given this setup, we consider selecting queries $a \in \mathcal{A}$ that approximately minimize

$$\max_{y\in\mathcal{Y}} \sum_{g,g'} \pi_t(g)\pi_t(g')d(g,g')e^{-\beta(\ell(g(a),y)+\ell(g'(a),y))}. \quad (5)$$

When $\ell(\cdot,\cdot)$ is the 0-1 loss and $\beta \to \infty$, the above corresponds to selecting queries that maximize average splitting. When $\mathcal{Y}$ is finite, we can still use SELECT to choose our query. However, we found that simply drawing a sequence of structure pairs and choosing the query that empirically minimizes equation (5) performed well enough.

Linear classifier simulations We consider the problem of learning linear classifiers where the data is distributed uniformly over the unit sphere $\mathcal{S}^{d-1}$. In this setting, there is a target classifier $w^* \in \mathbb{R}^d$, and the goal is to find a vector $w \in \mathbb{R}^d$ minimizing

$$d(w, w^*) = \Pr_{x\sim\mathrm{unif}(\mathcal{S}^{d-1})}(\mathrm{sign}(\langle w, x\rangle) \neq \mathrm{sign}(\langle w^*, x\rangle))$$

We ran experiments on actively learning such a classifier under the logistic noise model where $w^* \sim \mathcal{N}(0, \sigma^2 I_d)$ and $\Pr(y \,|\, x, w^*) = \left(1 + e^{-y\langle w^*, x\rangle}\right)^{-1}$.

Figure 1 shows the performance of NDBAL run with the logistic loss against two baselines: random sampling and QBC (Freund et al., 1997; Tosh and Dasgupta, 2018)–an active learner that repeatedly samples an atom and two structures and queries the atom if the two structures disagree on it.

Logit choice simulations In the logit choice model (Train, 2009), there is a fixed set of n items, represented as $x_1, \ldots, x_n \in \mathbb{R}^d$, and there is some consumer whose preferences over the items can be captured by a vector $w^* \in \mathbb{R}^d$, such that the consumer prefers item i over item j if and only if $\langle w^*, x_i\rangle > \langle w^*, x_j\rangle$. When presented with a pair of items (i, j), the consumer chooses item i with probability $1/(1 + e^{-\langle w^*, x_i - x_j\rangle})$.

We performed simulations in an interactive setting in which pairs of items are adaptively presented to the consumer. We considered two objectives.

(i) Best item identification: identifying $x_{i_{w^*}}$ where $i_w = \arg\max_i \langle w, x_i\rangle$ is the top item under w.

(ii) Approximate best item identification: finding an item j such that $\|x_j - x_{i_{w^*}}\|$ is small.

We generated $w^* \sim \mathcal{N}(0, \sigma^2 I_d)$ and drew $x_1, \ldots, x_n$ uniformly from S^{d-1}. To run NDBAL, we used $d(w, w') = \|x_{i_w} - x_{i_{w'}}\|$ as our structure distance. The results are displayed in Figure 1.

Experimental summary. In the appendix, we provide more settings of parameters as well as more information on our experimental setup. Across all our experiments, we found that NDBAL generally outperformed QBC and RANDOM on the metrics we tested.

Acknowledgements

The authors thank the anonymous reviewers for suggestions that improved the paper. They also acknowledge the NSF for support under grant CCF-1740833. CT also thanks Wesley Tansey for useful conversations that helped inspire this work.

References

D. Angluin and L. Valiant. Fast probabilistic algorithms for hamiltonian circuits and matchings. In *Proceedings of the 9th Annual ACM Symposium on Theory of Computing*, pages 30–41, 1977.

H. Ashtiani, S. Kushagra, and S. Ben-David. Clustering with same-cluster queries. In *Advances in Neural Information Processing Systems*, pages 3216–3224, 2016.

P. Awasthi, M.-F. Balcan, and K. Voevodski. Local algorithms for interactive clustering. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

P. Awasthi, M.-F. Balcan, N. Haghtalab, and R. Urner. Efficient learning of linear separators under bounded noise. In *Proceedings of the 28th Annual Conference on Learning Theory*, pages 167–190, 2015.

K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

M.-F. Balcan and S. Hanneke. Robust interactive learning. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. *Machine Learning*, 80(2-3):111–139, 2010.

J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. Margolin, S. Kim, C. Wilson, J. Lehár, G. Kryukov, and D. Sonkin. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.

S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*, 2005.

R. Dwivedi, Y. Chen, M.J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Proceedings of the 31st Annual Conference on Learning Theory*, pages 793–797, 2018.

E. Emamjomeh-Zadeh and D. Kempe. A general framework for robust interactive learning. In *Advances in Neural Information Processing Systems*, pages 7082–7091, 2017.

Y. Freund, H. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2):133–168, 1997.

J.B.S. Haldane. On a method of estimating frequencies. *Biometrika*, 33(3):222–225, 1945.

S. Hanneke. Rates of convergence in active learning. *Annals of Statistics*, 39(1):333–361, 2011.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

D. Haussler and P. M. Long. A generalization of Sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith. Interactive topic modeling. *Machine Learning*, 95: 423–469, 2014.

A. Krishnamurthy and A. Singh. On the power of adaptivity in matrix completion and approximation. *arXiv preprint arXiv:1407.3619*, 2014.

L. Lovasz and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30:307–358, 2007.

J. Lund, C. Cook, K. Seppi, and J. Boyd-Graber. Tandem anchoring: A multiword anchor approach for interactive topic modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 896–905, 2017.

S. Mohajer, C. Suh, and A. Elmahdy. Active learning for top-k rank aggregation from noisy comparisons. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2488–2497, 2017.

B.K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

R. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12): 7893–7906, 2011.

S. Resnick. *A probability path*. Springer Science & Business Media, 2013.

G.O. Roberts and J.S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

N. Shah and M. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

C. Tosh and S. Dasgupta. Diameter-based active learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3444–3452, 2017.

C. Tosh and S. Dasgupta. Interactive structure learning with structural query-by-committee. In *Advances in Neural Information Processing Systems*, 2018.

K. Train. *Discrete choice methods with simulation.* Cambridge university press, 2009.

S. Vikram and S. Dasgupta. Interactive Bayesian hierarchical clustering. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

W. Yang, J. Soares, P. Greninger, E. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. Smith, I.R. Thompson, et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961, 2012.