JMLR: Workshop and Conference Proceedings vol 23 (2012) 19.1–19.20 25th Annual Conference on Learning Theory

Active Learning Using Smooth Relative Regret Approximations with Applications

Nir Ailon[*] NAILON@CS.TECHNION.AC.IL
Dept. of Computer Science, Technion IIT, Haifa, Israel

Ron Begleiter RONBEG@CS.TECHNION.AC.IL
Dept. of Computer Science, Technion IIT, Haifa, Israel

Esther Ezra ESTHER@COURANT.NYU.EDU
Coutrant Institute of Maehtmatical Science, NYU, New York, NY

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

The disagreement coefficient of Hanneke has become a central concept in proving active learning rates. It has been shown in various ways that a concept class with low complexity together with a bound on the disagreement coefficient at an optimal solution allows active learning rates that are superior to passive learning ones.

We present a different tool for pool based active learning which follows from the existence of a certain uniform version of low disagreement coefficient, but is not equivalent to it. In fact, we present two fundamental active learning problems of significant interest for which our approach allows nontrivial active learning bounds. However, any general purpose method relying on the disagreement coefficient bounds only fails to guarantee any useful bounds for these problems.

The tool we use is based on the learner's ability to compute an estimator of the difference between the loss of any hypotheses and some fixed "pivotal" hypothesis to within an absolute error of at most ε times the ℓ_1 distance (the disagreement measure) between the two hypotheses. We prove that such an estimator implies the existence of a learning algorithm which, at each iteration, reduces its excess risk to within a constant factor. Each iteration replaces the current pivotal hypothesis with the minimizer of the estimated loss difference function with respect to the previous pivotal hypothesis. The label complexity essentially becomes that of computing this estimator.

The two applications of interest are: learning to rank from pairwise preferences, and clustering with side information (a.k.a. semi-supervised clustering). They are both fundamental, and have started receiving more attention from active learning theoreticians and practitioners.

Keywords: active learning, learning to rank from pairwise preferences, semi-supervised clustering, clustering with side information, disagreement coefficient, smooth relative regret approximation

1. Introduction

Unlike in standard PAC learning, an active learner chooses which instances to learn from. In the streaming setting, they may reject labels for instances arriving in a stream, and in the pool setting they may collect a pool of instances and then choose a subset from which to ask labels for. Although a relatively young field compared to traditional (passive) learning, there is by now a significant body of literature on the subject (see, e.g., Freund et al., 1997; Dasgupta, 2005; Castro et al., 2005; Kääriäinen, 2006; Balcan et al., 2006; Sugiyama, 2006; Hanneke, 2007; Balcan et al., 2007;

[*] Supported by a Marie Curie International Reintegration Grant PIRG07-GA-2010-268403

Dasgupta et al., 2007; Bach, 2007; Castro and Nowak, 2008; Balcan et al., 2008; Dasgupta and Hsu, 2008; Cavallanti et al., 2008; Hanneke, 2009; Beygelzimer et al., 2009, 2010; Koltchinskii, 2010; Cesa-Bianchi et al., 2010; Yang et al., 2010; Hanneke and Yang, 2010; El-Yaniv and Wiener, 2010; Hanneke, 2011; Orabona and Cesa-Bianchi, 2011; Cavallanti et al., 2011; Yang et al., 2011; Wang, 2011; Minsker, 2012). Refer to a survey by Settles (2009) for definition of active learning.

The disagreement coefficient of Hanneke (2007) has become a central data independent invariant in proving active learning rates. It has been shown in various ways that a concept class with low complexity together with a bound on the disagreement coefficient at an optimal solution allows active learning rates that are superior to passive rates under certain low noise conditions (see, e.g., Hanneke, 2007; Balcan et al., 2007; Dasgupta et al., 2007; Castro and Nowak, 2008; Beygelzimer et al., 2010). The best results assuming bounded VC dimension d and disagreement coefficient θ only can roughly be stated as follows: If the sought excess risk μ is the same order of magnitude as the optimal error ν or larger, then the number of required queries is roughly $\tilde{O}(\theta d \log(1/\mu))$.[1] Otherwise, the number is roughly $\tilde{O}(\theta d \nu^2/\mu^2)$. Note that these results make no assumption on the noise (except maybe for its magnitude). Better results can be made by assuming certain statistical properties of the noise (especially the model of Mammen and Tsybakov, 1999; Tsybakov, 2004).

The idea behind the disagreement coefficient is intuitive and simple. If a hypothesis h is r-close to optimal, then the *difference between their losses* (the regret of h) can be computed from instances in the *disagreement region* only, defined as the set of instances on which the r-ball round the optimal is not unanimous on. This means that for minimizing regret, one may restrict attention to hypotheses lying in iteratively shrinking *version spaces* and to instances in the corresponding disagreement region, which is shrinking in tandem with the version space if the disagreement coefficient is small. As pointed out by Beygelzimer et al. (2010), ignoring hypotheses outside the version space is brittle business, because a mistake in computation of the version space dooms the algorithm to fail. They propose a scheme in which no version space is computed. Instead, a certain importance weighted scheme is used. We also use importance weighting, but in the pool based setting and not in the streaming setting as they do.[2]

Analyzing the difference between losses ("relative regrets") of hypotheses is used almost in all theoretical work on active learning, but not attacked directly. In this work we argue that by carefully constructing empirical processes uniformly estimating the relative regret of all hypotheses with respect to a fixed "pivotal" hypothesis (the current solution) yields fast active learning rates. We call such constructions SRRA (Smooth Relative Regret Approximations).

We also show that (not surprisingly) low disagreement coefficient and VC dimension assumptions imply such efficient constructions, and give rise to yet another proof for the usefulness of the disagreement coefficient in active learning via an algorithm that does *not* need to compute or restrict itself to shrinking version spaces. We then present two fundamental pool based learning problems for which direct SRRA construction yields superior active learning rates, whereas any known argument that uses the disagreement coefficient only, requires the practitioner to obtain labels for the entire pool (!) even for moderately chosen parameters. We conclude that the SRRA method is, up to minor factors, at least as good as the disagreement coefficient method, but can be significantly better in certain cases.

1. The $\tilde{O}$ notation suppresses polylogarithmic terms.
2. Note that a practitioner can pretend that any pool based input is a stream, though that approach would probably not take full advantage of the data.

We note that another important line of design and analysis of active learning algorithms makes certain structural or Vayesian assumptions on the noise (e.g., Balcan et al., 2007; Castro and Nowak, 2008; Hanneke, 2009; Koltchinskii, 2010; Yang et al., 2010; Wang, 2011; Yang et al., 2011; Minsker, 2012). We expect that one can get yet improved analysis in our framework under these assumptions. We leave this to future work.

The rest of the paper is laid out as follows: In Section 2 we present notations and basic definitions, including an introduction to our method. In Section 3 we show that existence of low disagreement coefficient implies our method, in some sense. In Section 4 we present our two main applications, learning to rank from pairwise preferences (LRPP) in Section 4.1 and clustering with side information in Section 4.2. Finally in Section 5 we present additional results and practical considerations, and in particular how to use our methods with convex relaxations if the ERM[3] problems that arise in the discussion are too difficult (computationally) to optimally solve. We conclude in Section 6 and suggest future directions. Due to lack of space, all proofs, as well as certain literature surveys and historical notes appear in a full version text (Ailon et al., 2012).

2. Definitions and Notation

We follow the notation of Hanneke (2011): Let $\mathcal{X}$ be an instance space, and let $\mathcal{Y} = \{0, 1\}$ be a label space. Denote by $\mathcal{D}$ the distribution over $\mathcal{X} \times \mathcal{Y}$, with corresponding marginals $\mathcal{D}_{\mathcal{X}}$ and $\mathcal{D}_{\mathcal{Y}}$. In this work we assume for convenience that each label Y is a deterministic function of X, so that if $X \sim \mathcal{D}_X$ then $(X, Y(X))$ is distributed according to $\mathcal{D}$.

By $\mathcal{C}$ we denote a concept class of functions mapping $\mathcal{X}$ to $\mathcal{Y}$. The error rate of a hypothesis $h \in \mathcal{C}$ equals

$$\mathrm{er}_{\mathcal{D}}(h) = E_{(X,Y)\sim\mathcal{D}}[h(X) \neq Y] \,.$$

The noise rate ν of $\mathcal{C}$ is defined as $\nu = \inf_{h\in\mathcal{C}} \mathrm{er}_{\mathcal{D}}(h)$. We will focus on the scenario in which ν is attained at an optimal hypothesis h^*, so that $\mathrm{er}_{\mathcal{D}}(h^*) = \nu$. Define the distance $\mathrm{dist}(h_1, h_2)$ between two hypotheses $h_1, h_2 \in C$ as $\Pr_{X\sim\mathcal{D}_{\mathcal{X}}}[h_1(X) \neq h_2(X)]$; observe that $\mathrm{dist}(\cdot, \cdot)$ is a pseudo-metric over pairs of hypotheses. For a hypothesis $h \in \mathcal{C}$ and a number $r \geq 0$, the ball $\mathcal{B}(h, r)$ around h of radius r is defined as $\{h' \in \mathcal{C} : \ \mathrm{dist}(h, h') \leq r\}$. For a set $V \subseteq \mathcal{C}$ of hypotheses, let $\mathrm{DIS}(V)$ denote

$$\mathrm{DIS}(V) = \{x \in \mathcal{X} : \exists h_1, h_2 \in V \text{ s.t. } h_1(x) \neq h_2(x)\} \,.$$

2.1. The Disagreement Coefficient

The disagreement coefficient of h with respect to $\mathcal{C}$ under $\mathcal{D}_{\mathcal{X}}$ is defined as

$$\theta_h = \sup_{r>0} \frac{\Pr_{\mathcal{D}_{\mathcal{X}}}[\mathrm{DIS}(\mathcal{B}(h, r))]}{r} \,, \tag{2.1}$$

where $\Pr_{\mathcal{D}_{\mathcal{X}}}[\mathcal{W}]$ for $\mathcal{W} \subseteq \mathcal{X}$ denotes the probability measure with respect to the distribution $\mathcal{D}_{\mathcal{X}}$. Define the uniform disagreement coefficient θ as $\sup_{h\in\mathcal{C}} \theta_h$, namely

$$\theta = \sup_{h\in\mathcal{C}} \sup_{r>0} \frac{\Pr_{\mathcal{D}_{\mathcal{X}}}[\mathrm{DIS}(\mathcal{B}(h, r))]}{r} \,. \tag{2.2}$$

3. Empirical Risk Minimization.

Remark 1 *A useful slight variation of the definitions of θ_h and θ can be obtained by replacing* $\sup_{r>0}$ *with* $\sup_{r\geq\nu}$ *in (2.1) and (2.2). We will explicitly say when we refer to this variation in what follows.*

2.2. Smooth Relative Regret Approximations (SRRA)

Fix $h \in \mathcal{C}$ (which we call the *pivotal hypothesis*). Denote by $\mathrm{reg}_h : \mathcal{C} \mapsto \mathbf{R}$ the function defined as

$$\mathrm{reg}_h(h') = \mathrm{er}_{\mathcal{D}}(h') - \mathrm{er}_{\mathcal{D}}(h) \ .$$

We call reg_h the *relative regret function with respect to* h. Note that for $h = h^*$ this is simply the usual regret, or excess risk function.

Definition 2 *Let $f : \mathcal{C} \mapsto \mathbf{R}$ be any function, and $0 < \varepsilon < 1/5$ and $0 < \mu \leq 1$. We say that f is an* (ε, μ)-smooth relative regret approximation ((ε, μ)-SRRA) *with respect to h if for all $h' \in \mathcal{C}$,*

$$|f(h') - \mathrm{reg}_h(h')| \leq \varepsilon \cdot \left(\mathrm{dist}(h, h') + \mu\right) \ .$$

If $\mu = 0$ we simply say that f is an ε-smooth relative regret approximation *with respect to h.*

Although the definition is general, the applications we study in details fall under the category of pool based active learning, in which $\mathcal{X}$ is a finite set and $\Pr_{\mathcal{D}_{\mathcal{X}}}$ is the uniform measure. This allow us to take $\mu = 0$, and will be useful in what follows. The following theorem and corollary constitute the main ingredient in our work. The corresponding proofs are deferred to Appendices A, and B.

Theorem 3 *Let $h \in \mathcal{C}$ and f be an (ε, μ)-SRRA with respect to h. Let $h_1 = \mathrm{argmin}_{h' \in \mathcal{C}} f(h')$. Then*

$$\mathrm{er}_{\mathcal{D}}(h_1) = (1 + O(\varepsilon))\, \nu + O\left(\varepsilon \cdot \mathrm{er}_{\mathcal{D}}(h)\right) + O(\varepsilon\mu) \ .$$

A simple inductive use of the theorem proves the following corollary, bounding the query complexity of an ERM based active learning algorithm (see Algorithm 1 for corresponding pseudocode). Note that the proposed algorithm never restricts itself to a shrinking version space.

Corollary 4 *Let $h_0, h_1, h_2, \ldots$ be a sequence of hypotheses in $\mathcal{C}$ such that for all $i \geq 1$, $h_i = \mathrm{argmin}_{h' \in \mathcal{C}} f_{i-1}(h')$, where f_{i-1} is an (ε, μ)-SRRA with respect to h_{i-1}. Then for all $i \geq 0$,*

$$\mathrm{er}_{\mathcal{D}}(h_i) = (1 + O(\varepsilon))\, \nu + O(\varepsilon^i)\mathrm{er}_{\mathcal{D}}(h_0) + O(\varepsilon\mu) \ .$$

We will show below problems of interest in which (ε, μ)-SRRA's with respect to a given hypothesis h can be obtained using queries $Y(X)$ at few randomly (and adaptively) selected points $X \in \mathcal{X}$, if the uniform disagreement coefficient θ is small. This will constitute another proof for the usefulness of the disagreement coefficient in design and analysis of active learning algorithms. We then present two problems for which a direct construction of an SRRA yields a significantly better query complexity than that guaranteed using the disagreement coefficient alone.

3. Constant Uniform Disagreement Coefficient Implies Efficient SRRA's

We show that a bounded uniform disagreement coefficient implies existence of query efficient (ε, μ)-SRRA's. This constitutes yet another proof of the usefulness of the disagreement coefficient in design of active learning algorithms, via Algorithm 1.

Algorithm 1 An Active Learning Algorithm from SRRA's

Input: an initial solution $h_0 \in \mathcal{C}$, estimation parameters $\epsilon \in (0, 1/5), \mu > 0$, and number of iterations T

1: $i \leftarrow 0$
2: **repeat**
3: $h_{i+1} \leftarrow \operatorname{argmin}_{h' \in \mathcal{C}} f(h')$, where f is an (ε, μ)-smooth relative regret approximation with respect to h_i
4: $i \leftarrow i + 1$
5: **until** i equals T
6: **return** h_T

3.1. The Construction

Returning to our problem, assume the uniform disagreement coefficient θ corresponding to $\mathcal{C}$ is finite and $\nu > 0$. Fix some failure probability δ. We consider the range space $(\mathcal{X}, \mathcal{C}^*)$, defined by

$$\mathcal{C}^* = \left(\bigcup_{h' \in \mathcal{C}} \left\{ \{X \in \mathcal{X} : h'(X) = 0\} \right\} \right) \cup \left(\bigcup_{h' \in \mathcal{C}} \left\{ \{X \in \mathcal{X} : h'(X) = 1\} \right\} \right) .$$

In other words, $\mathcal{C}^*$ is the collection of all subsets $S \subseteq \mathcal{X}$, whose elements $X \in S$ are mapped to the same value (0 or 1) by h', for some $h' \in \mathcal{C}$. Assume $(\mathcal{X}, \mathcal{C}^*)$ has VC dimension d, and fix $h \in \mathcal{C}$. Let $L = \lceil \log \mu^{-1} \rceil$. Define $\mathcal{X}_0 = \mathrm{DIS}(\mathcal{B}(h, \mu))$ and for $i = 1, 2, \ldots, L$, define $\mathcal{X}_i$ to be

$$\mathcal{X}_i = \mathrm{DIS}(\mathcal{B}(h, \mu 2^i)) \setminus \mathrm{DIS}(\mathcal{B}(h, \mu 2^{i-1})) .$$

Let $\eta_i = \Pr_{\mathcal{D}_\mathcal{X}}[\mathcal{X}_i]$ be the measure of $\mathcal{X}_i$. For each $i \geq 0$ draw a sample $X_{i,1}, \ldots, X_{i,m}$ of $m = O\left(\varepsilon^{-2}\theta\left(d \log \theta + \log\left(\delta^{-1} \log(1/\mu)\right)\right)\right)$ examples in $\mathcal{X}_i$, each of which drawn independently from the distribution $\mathcal{D}_\mathcal{X}|\mathcal{X}_i$ (with repetitions). (By $\mathcal{D}_\mathcal{X}|\mathcal{X}_i$ we mean, the distribution $\mathcal{D}_\mathcal{X}$ conditioned on $\mathcal{X}_i$.) We will now define an estimator function $f : \mathcal{C} \mapsto \mathbf{R}$ of reg_h, as follows. For any $h' \in \mathcal{C}$ and $i = 0, 1, \ldots, L$ let

$$f_i(h') \triangleq \eta_i m^{-1} \sum_{j=1}^{m} \left(\mathbf{1}_{Y(X_{i,j}) \neq h'(X_{i,j})} - \mathbf{1}_{Y(X_{i,j}) \neq h(X_{i,j})} \right) .$$

Our estimator is now defined as $f(h') \triangleq \sum_{i=0}^{L} f_i(h')$. We next show:

Theorem 5 *Let f, h, h', m be as above. With probability at least $1 - \delta$, f is an (ε, μ)-SRRA with respect to h.*

A main tool to be exploited in the proof is called *relative ε-approximations*. We defer the details to Appendix C (see also Li et al., 2000, for further details). We conclude:

Corollary 6 *An (ε, μ)-SRRA with respect to h can be constructed, with probability at least $1 - \delta$, using at most*

$$m\left(1 + \lceil \log(1/\mu) \rceil\right) = O\left(\theta\varepsilon^{-2}(\log(1/\mu))\left(d \log \theta + \log(\delta^{-1} \log(1/\mu))\right)\right) \tag{3.1}$$

label queries.

Combining Corollaries 4 and 6 (Algorithm 1), we obtain an active learning algorithm in the ERM setting, with query complexity depending on the uniform disagreement coefficient and the VC dimension. Assume δ is a constant. If we are interested in excess risk of $c\nu$ for some constant constant $c > 0$, then we may take ε to be $\theta(c)$ and $\mu = \theta(\nu)$ and build an (ε, μ)-SRRA's using $O(\theta d(\log(1/\mu))(\log\theta))$, once for each of $O(\log(1/\nu))$ iterations of Algorithm 1. If we seek a solution with error $(1+\varepsilon)\nu$, we would need to construct (ε, ν)-SRRA's using $O(\theta d\varepsilon^{-2}(\log(1/\nu))(\log\theta))$ query labels, one for each of $O(\log(1/\nu))$ iterations of the algorithm. The total label query complexity for a fixed ε value is $O(\theta d(\log^2(1/\mu))(\log\theta))$, which is $O(\log(1/\nu))$ times the best known bounds using disagreement coefficient and VC dimension bounds only.

A few more comparison notes are in place. First, note that in known arguments bounding query complexity using the disagreement coefficient, the disagreement coefficient θ_{h^*} with respect to the optimal hypothesis h^* is used in the analysis, and not the uniform coefficient θ. Also note that both in previously known results bounding query complexity using disagreement coefficient and VC dimension bounds as well as in our result, the slight improvement described in Remark 1 applies. That is, all arguments remain valid if we replace the supremums in (2.1) and (2.2) with $\sup_{r\geq\nu}$.

4. Two Important Applications

In this section, we present two cases for which we construct (ε, μ)-SRRA's directly, and thus obtain query efficient active learning algorithms. On the flip side, we show that any known argument based on the disagreement coefficient and the VC dimension only, and in particular Corollary 6 fails to guarantee useful active learning bounds. From this we conclude that, although bounded disagreement coefficient and VC dimension may lead to construction of SRRA's, studying SRRA's directly may allow stronger query complexity bounds.

The setting of these two problems is basically a distribution-free setting over finite $\mathcal{X}$ (that is "transductive"). Taking $\Pr_{\mathcal{D}_\mathcal{X}}$ to be the uniform measure allows us to keep with the original definitions of Section 2. In addition it allows us to take $\mu = 0$. Thus, instead of using $(\varepsilon, 0)$-SRRA we simply ignore the parameter μ and refer to our estimators as ε-SRRA in what follows.

4.1. Application #1: Learning to Rank from Pairwise Preferences (LRPP)

We describe a learning problem of significant interest. We refer the reader to Appendix D for a detailed history of the problem, which we omit due to lack of space, except for necessary direct comparisons to previous work, which we mention as we go.

Let V be a set of n elements (alternatives). The instance space $\mathcal{X}$ is taken to be the set of all distinct pairs of elements in V, namely $V \times V \setminus \big\{(u,u) : u \in V\big\}$. The distribution $\mathcal{D}_\mathcal{X}$ is uniform on $\mathcal{X}$ The label function $Y : \mathcal{X} \mapsto \{0,1\}$ encodes a preference function satisfying $Y\big((u,v)\big) = 1 - Y\big((v,u)\big)$ for all $u, v \in V$.[4] By convention, we think of $Y\big((u,v)\big) = 1$ as a stipulation that u is preferred over v. For convenience we will drop the double-parentheses in what follows.

The class of solution functions $\mathcal{C}$ we consider is all $h : \mathcal{X} \to \{0,1\}$ such that $h(u,v) = 1 - h(v,u)$ and $h(u,z) \leq h(u,v) + h(v,z)$ (transitivity) for all distinct $u, v, z \in V$. This is

4. We chose this definition for convenience in what follows. Note, however, that we could have defined $\mathcal{X}$ to be unordered pairs of elements in V without making any assumptions on Y. For example, by assuming an arbitrary indexing order over $\mathcal{X}$ and thinking of $Y(\{u,v\}) = 1$ as a stipulation that the alternative with minimal index is preferred over its counterpart.

equivalent to the space of permutations over V, and we will use the notation $\pi, \sigma, \ldots$ instead of $h, h', \ldots$ in the remainder of the section. We also use notation $u \prec_\pi v$ as a predicate equivalent to $\pi(u,v) = 1$. Endowing $\mathcal{X}$ with the uniform measure, $\mathrm{dist}(\pi,\sigma)$ turns out to be (up to normalization) the well known kendall-τ distance: $\mathrm{dist}(\pi,\sigma) = N^{-1}\sum_{u\neq v} \mathbf{1}_{\pi(u,v)\neq\sigma(u,v)}$, where $N \triangleq n(n-1)$ is the number of all ordered pairs.

Let us first see if we can get a useful active learning algorithm using disagreement coefficient arguments. It has been shown in Ailon (2012) that the uniform disagreement coefficient of $\mathcal{C}$ is $\Theta(n)$ (to see this simple fact, notice that if we start from some permutation π and swap the positions of *any* two elements $u, v \in V$, then we obtain a permutation of distance at most $O(1/n)$ away from π, hence the disagreement region of the ball of radius $O(1/n)$ around π is the entire space $\mathcal{X}$). It is also known that the VC dimension of $\mathcal{C}$ is $n-1$ (see, Radinsky and Ailon, 2011). Using Corollary 6, we conclude that we would need $\Omega(n^2)$ preference labels to obtain an (ε,μ)-SRRA for any meaningful pair (ε,μ). This is uninformative because the cardinality of $\mathcal{X}$ is $O(n^2)$. A similar bound is obtained using any known active learning bound using disagreement coefficient and VC-dimension bounds only.

Remark: A slight improvement can be obtained using Remark 1: Using the refined definition of disagreement coefficient, it is not hard to see that the uniform disagreement coefficient, as well as the disagreement coefficient at the optimal solution h^* becomes[5] $\theta = \theta_{h^*} = O(1/\nu)$, if $\nu \geq n^{-1}$. This improves the query complexity bound to $O(n\nu^{-1})$. If ν tends to n^{-1} from above, in the limit this becomes a quadratic (in n) query complexity.

We next show how to construct more useful (in terms of query complexity) SRRA's for LRPP, *for arbitrarily small* ν.

4.1.1. BETTER SRRA FOR LRPP

Consider the following idea for creating an ε-SRRA for LRPP, with respect to some fixed $\pi \in \mathcal{C}$. We start by defining the following sample size parameter:

$$p \triangleq O\left(\varepsilon^{-3}\log^3 n\right) \ . \tag{4.1}$$

For all $u \in V$ and for all $i = 0, 1, \ldots, \lceil \log n \rceil$, let $I_{u,i}$ denote the set of elements v such that $2^i p \leq |\pi(u) - \pi(v)| < 2^{i+1}p$ where, abusing notation, $\pi(u)$ is the position of u in π (e.g. $\pi(u)$ is 1 if u beats all other elements, and n if it is beaten by all others). From this set, choose a random subset $R_{u,i}$ of $\lceil |I_{u,i}|/2^i \rceil$ elements, each chosen uniformly (with repetitions).[6,7] For distinct $u, v \in V$ and a permutation $\sigma \in \mathcal{C}$, let $C_{u,v}(\sigma)$ denote the contribution of the pair u, v to $\mathrm{er}_\mathcal{D}(\sigma)$, namely:

$$C_{u,v}(\sigma) \triangleq N^{-1}\mathbf{1}_{\sigma(u,v)\neq Y(u,v)} \ . \tag{4.2}$$

(Note that $C_{u,v} \equiv C_{v,u}$.) Our estimator $f(\sigma)$ of $\mathrm{reg}_\pi(\sigma) = \mathrm{er}_\mathcal{D}(\sigma) - \mathrm{er}_\mathcal{D}(\pi)$ is defined as

$$f(\sigma) = \sum_{u\in V}\sum_{i=0}^{\lceil \log n\rceil} \frac{|I_{u,i}|}{|R_{u,i}|}\sum_{v\in R_{u,i}}\left(C_{u,v}(\sigma) - C_{u,v}(\pi)\right) \ . \tag{4.3}$$

5. Due to symmetry, the uniform disagreement coefficient here equals θ_h for any $h \in \mathcal{C}$.
6. Think of the size of the set $R_{u,i}$ as exactly p - the number $\lceil |I_{u,i}|/2^i \rceil$ is a technicality required for dealing with sets $I_{u,i}$ that are clipped at the edges and are not of maximal size $2^i p$.
7. A variant of this sampling scheme is as follows: For each pair (u,v), add it to S with probability proportional to $\min\{1, p/|\pi(u) - \pi(v)|\}$. A similar scheme can be found in the work of Ailon et al. (2007); Halevy and Kushilevitz (2007); Ailon (2012), but the strong properties proven here were not known.

Note that the inner sum treats $R_{u,i}$ as a multi-set, because elements were chosen with repetition. Clearly, $f(\sigma)$ is an unbiased estimator of $\mathrm{reg}_\pi(\sigma)$. Our goal is to prove that $f(\sigma)$ is an ε-SRRA.

Theorem 7 *With probability at least $1 - n^{-3}$, the function f is an ε-SRRA with respect to π.*

The proof, which is deferred to the full version (Ailon et al., 2012), is based on decomposing $\mathrm{reg}_h(h') - f(h')$ and $\mathrm{dist}(h', h)$ vis a vis a careful partition of $\mathcal{X}$. Note that the number of preference queries required for computing f is $O(\varepsilon^{-3} n \log^3 n)$. We conclude from the theorem, our bound on the number of preference queries and the iterative algorithm described in Corollary 4 (see Algorithm 1):

Corollary 8 *There exists an active learning algorithm for obtaining a solution $\pi \in \mathcal{C}$ for LRPP with $\mathrm{er}_{\mathcal{D}}(\pi) \leq (1 + O(\varepsilon))\,\nu$ with total query complexity of $O\left(\varepsilon^{-3} n \log^4 n\right)$. The algorithm succeeds with probability at least $1 - n^{-2}$.*

The corollary improves a recent result (Ailon, 2012) in two ways. First, it shaves off $\log n$ and ε^{-1} factors from the query complexity. Secondly, and more importantly, by using Algorithm 1, our optimization method avoids the divide and conquer method on which Ailon (2012) heavily relied on, and thus lifts a highly restrictive practical requirement arising when searching in *restricted permutation spaces*. We refer the reader to Appendix D for a more detailed explanation.

Corollary 8 allows us to find a solution of cost $(1 + \varepsilon)\nu$ with query complexity that is slightly above linear in n (for constant ε), regardless of the magnitude of ν. In comparison, as we saw in Section 4.1, known active learning results (and in particular Corollary 6) that used disagreement coefficient and VC dimension bounds only guaranteed a query complexity of $\Omega(n\nu^{-1})$, tending to the pool size of $n(n-1)$ as ν becomes small. Note that $\nu = o(1)$ is quite realistic for this problem. For example, consider the following noise model. A ground truth permutation π^* exists, $Y(u, v)$ is obtained as a human response to the question of preference between u and v with respect to π^*, and the human errs with probability proportional to $|\pi^*(u) - \pi^*(v)|^{-\rho}$, for some $\rho > 0$. Namely, closer pairs of item in the ground truth permutation are more prone to confuse a human labeler. This is quite natural. The resulting noise is $\nu = n^{-\rho}$.[8]

4.2. Application #2: Clustering with Side Information

We refer the reader again to Appendix D for a detailed history and account of previous results for the problem below, which we omit for lack of space.

Let V be a set of points of size n. Our goal now is to partition V into k sets (clusters), where k is fixed. In most applications, V is endowed with some metric, and the practitioner uses this metric in order to evaluate the quality of a clustering solution. In some cases, known as *semi-supervised clustering*, or *clustering with side information*, additional information comes in the form of *pairwise constraints*. Such a constraint tells us for a pair $u, v \in V$ whether they should be in the same cluster or in separate ones. We concentrate on using such information.

Using the notation of our framework, $\mathcal{X}$ denotes the set of distinct pairs of elements in V (same as in Section 4.1), and $\mathcal{D}_{\mathcal{X}}$ is the corresponding uniform measure. The label $Y\big((u, v)\big) = 1$ means that u and v should be clustered together, and $Y\big((u, v)\big) = 0$ means the opposite. Assume that $Y\big((u, v)\big) = Y\big((v, u)\big)$ for all u, v.[9]

8. Our work does not assume Bayesian noise, and we present this scenario for illustration purposes only.
9. Equivalently, assume that $\mathcal{X}$ contains only unordered distinct pairs without any constraint on Y. For notational purposes we preferred to define $\mathcal{X}$ as the set of ordered distinct pairs.

The concept class $\mathcal{C}$ is the set of equivalence relations over V with at most k equivalence classes. More precisely, every $h \in \mathcal{C}$ is identified with a disjoint cover $V_1, \ldots, V_k$ of V (some V_i's possible empty), with $h\big((u,v)\big) = 1$ if and only if $u, v \in V_j$ for some j. As usual, Y may induce a non-transitive relation (e.g., we could have $Y\big((u,v)\big) = Y\big((v,z)\big) = 1$ and $Y\big((u,z)\big) = 0$). In what follows, we will drop the double parentheses. Also, we will abuse notation by viewing h as both an equivalence relation and as a disjoint cover $\{V_1, \ldots, V_k\}$ of V. We take $\mathcal{D}_{\mathcal{X}}$ to be the uniform measure on $\mathcal{X}$. The error of $h \in \mathcal{C}$ is given as $\mathrm{er}_{\mathcal{D}}(h) = N^{-1} \sum_{(u,v)\in\mathcal{X}} \mathbf{1}_{h(u,v)\neq Y(u,v)}$ where, as before, $N = |\mathcal{X}| = n(n-1)$. We will define $\mathrm{cost}_{u,v}(h)$ to be the contribution $N^{-1}\mathbf{1}_{h(u,v)\neq Y(u,v)}$ of $(u,v) \in \mathcal{X}$ to $\mathrm{er}_{\mathcal{D}}$. The distance $\mathrm{dist}(h,h')$ is given as $\mathrm{dist}(h,h') = N^{-1} \sum_{(u,v)\in\mathcal{X}} \mathbf{1}_{h(u,v)\neq h'(u,v)}$.

We check again what disagreement coefficient based arguments can contribute to this problem. It is easy to see that the uniform disagreement coefficient of $\mathcal{C}$ is $\Theta(n)$. Indeed, starting from any solution $h \in \mathcal{C}$ with corresponding partitioning $\{V_1, \ldots, V_k\}$, consider the partition obtained by moving an element $u \in V$ from its current part V_j to some other part $V_{j'}$ for $j' \neq j$. In other words, consider the clustering $h' \in \mathcal{C}$ given by $\big\{V_{j'} \cup \{u\}, V_j \setminus \{u\}\big\} \cup \bigcup_{i\notin\{j,j'\}} \{V_i\}$. Observe that $\mathrm{dist}(h,h') = O(1/n)$. On the other hand, for any $v \in V$ and for any $u \in V$ there is a choice of j' so that h and h' obtained as above would disagree on $(u,v) \in \mathcal{X}$. Hence, $\Pr_{\mathcal{D}_{\mathcal{X}}}\left[\mathrm{DIS}\left(\mathcal{B}\left(h, O(1/n)\right)\right)\right] = 1$.

It is also not hard to see that the VC dimension of $\mathcal{C}$ is $\Theta(n)$. Indeed, any full matching over V constitutes a set which is shattered in $\mathcal{C}$ (as long as $k \geq 2$, of course). On the other hand, any set $S \subseteq \mathcal{X}$ of size n must induce an undirected cycle on the elements of V. Clearly the edges of a cycle cannot be shattered by functions in $\mathcal{C}$, because if $h(u_1,u_2) = h(u_2,u_3) = \cdots = h(u_{\ell-1},u_\ell) = 1$ for $h \in \mathcal{C}$, then also $h(u_1,u_\ell) = 1$.

Using Corollary 6, we conclude that we'd need $\Omega(n^2)$ preference labels to obtain an (ε,μ)-SRRA for any meaningful pair (ε,μ). This is uninformative because the cardinality of $\mathcal{X}$ is $O(n^2)$. As in the problem discussed in Section 4.1, this can be improved using Remark 4.1 to $\Omega(n\nu^{-1})$, which tends to quadratic in n as ν becomes smaller. We next show how to construct more useful SRRA's for the problem, for arbitrarily small ν.

4.2.1. BETTER SRRA FOR SEMI-SUPERVISED k-CLUSTERING

Fix $h \in \mathcal{C}$, with $h = \{V_1, \ldots, V_k\}$ (we allow empty V_i's). Order the V_i's so that $|V_1| \geq |V_2| \geq \cdots \geq |V_k|$. We construct an ε-SRRA with respect to h as follows. For each cluster $V_i \in h$ and for each element $u \in V_i$ we draw $k-i+1$ independent samples $S_{ui}, S_{u(i+1)}, \ldots, S_{uk}$ as follows. Each sample S_{uj} is a subset of V_j of size q (to be defined below), chosen uniformly with repetitions from V_j, where

$$q = c_2 \max\left\{\varepsilon^{-2}k^2, \varepsilon^{-3}k\right\} \log n \tag{4.4}$$

for some global $c_2 > 0$. Note that the collection of pairs $\{(u,v) \in \mathcal{X} : v \in S_{ui} \text{ for some } i\}$ is, roughly speaking, biased in such a way that pairs containing elements in smaller clusters (with respect to h) are more likely to be selected.

We define our estimator f to be, for any $h' \in \mathcal{C}$,

$$f(h') = \sum_{i=1}^{k} \frac{|V_i|}{q} \sum_{u\in V_i} \sum_{v\in S_{ui}} f_{u,v}(h') + 2\sum_{i=1}^{k} \sum_{u\in V_i} \sum_{j=i+1}^{k} \frac{|V_j|}{q} \sum_{v\in S_{uj}} f_{u,v}(h')\,, \tag{4.5}$$

where $f_{u,v}(h') \triangleq \mathrm{cost}_{u,v}(h') - \mathrm{cost}_{u,v}(h)$ and $\mathrm{cost}_{u,v}(h) \triangleq N^{-1}\mathbf{1}_{h(u,v)\neq Y(u,v)}$. Note that the summations over S_{ui} above takes into account multiplicity of elements in the multiset S_{ui}.

Theorem 9 *With probability at least* $1 - n^{-3}$ *the function* f *is an* ε*-SRRA with respect to* h*.*

The proof, which is deferred to the full version (Ailon et al., 2012), is based on decomposing $\mathrm{reg}_h(h') - f(h')$ and $\mathrm{dist}(h', h)$ vis a vis a careful partition of $\mathcal{X}$. The technical details are quite involved and require a non intuitive charging scheme. Clearly the number of label queries required for obtaining the ε-SRRA is $O(n \max\{\varepsilon^{-2}k^3, \varepsilon^{-3}k^2\} \log n)$. Combining the theorem with this bound and the iterative algorithm described in Corollary 4 (Algorithm 1), we obtain the following:

Corollary 10 *There exists an active learning algorithm for obtaining a solution* $\pi \in \mathcal{C}$ *for LRPP with* $\mathrm{er}_{\mathcal{D}}(\pi) \leq (1 + O(\varepsilon))\,\nu$ *with total query complexity of* $O\left(n \max\left\{\varepsilon^{-2}k^3, \varepsilon^{-3}k^2\right\} \log^2 n\right)$*. The algorithm succeeds with success probability at least* $1 - n^{-2}$*.*

As in the case of Corollary 8 and the ensuing discussion around LRPP, this significantly beats known active learning results using disagreement coefficient and VC dimension bounds only, for small ν.

5. Additional Results and Practical Considerations

We discuss two practical extensions of our results.

5.1. LRPP over Linearly Induced Permutations in Constant Dimensional Feature Space

A special class of interest is known as LRPP over linearly induced permutations in constant dimensional feature space. We use the same definition of $\mathcal{X}$ as in Section 4.1, except that now each point $v \in V$ is associated with a feature vector, which we denote using bold face: $\mathbf{v} \in \mathbf{R}^d$. The concept space $\mathcal{C}$ now consists only of permutations π such that there exists a vector $\mathbf{w}_\pi \in \mathbf{R}^d$ satisfying

$$\pi(u, v) = 1 \iff \langle \mathbf{w}, \mathbf{u} - \mathbf{v} \rangle > 0\,. \tag{5.1}$$

We are assuming familiarity with the theory of geometric arrangements, and refer the reader to de Berg et al. (2008) for further details. Geometrically, each $(u, v) \in \mathcal{X}$ is viewed as a halfspace $H_{u,v} \triangleq \{\mathbf{x} : \langle \mathbf{x}, \mathbf{u} - \mathbf{v} \rangle > 0\}$, whose (closure) supporting hyperplane is $h_{u,v} \triangleq \{\mathbf{x} : \langle \mathbf{x}, \mathbf{u} - \mathbf{v} \rangle = 0\}$. Let $\mathcal{H}$ be the collection of these $\binom{n}{2}$ hyperplanes $\{h_{u,v} : (u, v) \in \mathcal{X}\}$.[10] The collection $\mathcal{C}$ corresponds to the maximal dimensional cells in the underlying arrangement $\mathcal{A}(\mathcal{H})$. We thus call $\mathcal{A}(\mathcal{H})$ from now on the *permutation arrangement*, and we naturally identify full dimensional cells with their induced permutations. We denote by $\mathbb{C}_\pi \subseteq \mathbf{R}^d$ the unique cell corresponding to a permutation $\pi \in \mathcal{C}$.

Bounding the VC dimension and disagreement coefficient. Using standard tools from combinatorial geometry, the VC dimension of $\mathcal{C}$ is at most $d - 1$. Roughly speaking, this property follows from the fact that in an arrangement of m hyperplanes in d-space, each of which meeting the origin, the overall number of cells is at most $O(m^{d-1})$, see de Berg et al. (2008).

As for the uniform disagreement coefficient, we show below that it is bounded by $O(n)$. Let $\pi \in \mathcal{C}$ be a permutation with a corresponding cell $\mathbb{C}_\pi$ in $\mathcal{A}(\mathcal{H})$. The ball $\mathcal{B}(\pi, r)$ is, geometrically, the closure of the union of all cells corresponding to "realizable" permutations σ satisfying $\mathrm{dist}(\sigma, \pi) \leq r$. The corresponding disagreement region $\mathrm{DIS}(\mathcal{B}(\pi, r))$ corresponds to the set of ordered pairs (halfspaces) intersecting $\mathcal{B}(\pi, r)$. We next show:

10. Note that $h_{u,v} = h_{v,u}$.

Proposition 11 *The measure of* $\mathrm{DIS}\left(\mathcal{B}(\pi, r)\right)$ *in* $\mathcal{D}_{\mathcal{X}}$ *is at most* $8rn$.

The proof is deferred to the full version (Ailon et al., 2012). By the proposition we have that the disagreement coefficient θ is always bounded by $O(n)$, establishing our bound. We now invoke Corollary 6 with $\mu = O(1/n^2)$ (which is tantamount to $\mu = 0$ for this problem, because $|\mathcal{X}| = O(n^2)$ and we are using the uniform measure), and conclude:

Theorem 12 *An* ε*-SRRA for LRPP in linearly induced permutations in* d *dimensional feature space can be constructed, with respect to any* $\pi \in \mathcal{C}$*, with probability at least* $1 - \delta$*, using at most* $O\left(nd\varepsilon^{-2}\log^2 n + n\varepsilon^{-2}(\log n)\left(\log(\delta^{-1}\log n)\right)\right)$ *label queries.*

Combining Theorem 12, and the iterative algorithm described in Corollary 4:

Corollary 13 *There exists an algorithm for obtaining a solution* $\pi \in \mathcal{C}$ *for LRPP in linearly induced permutations in* d *dimensional feature space with* $\mathrm{er}_{\mathcal{D}}(\pi) \leq (1 + O(\varepsilon))\,\nu$ *with total query complexity of*

$$O\Big(\varepsilon^{-2}nd\log^3 n + n\varepsilon^{-2}(\log^2 n)\left(\log(\delta^{-1}\log n)\right)\Big) \tag{5.2}$$

The algorithm succeeds with success probability at least $1 - \delta$.

We compare this bound to that of Corollary 8. For the sake of comparison, assume $\delta = n^{-2}$, so that (5.2) takes the simpler form of $O\big(\varepsilon^{-2}nd\log^3 n\big)$. This bound is better than that of Corollary 8 as long as the feature space dimension d is $O(\varepsilon^{-1}\log n)$. For larger dimensions, Corollary 8 gives a better bound. It would be interesting to obtain a smoother interpolation between the *geometric* structure coming from the feature space and the *combinatorial* structure coming from permutations. We refer the reader to (Jamieson and Nowak, 2011) for a recent result with improved query complexity under certain Bayesian noise assumptions.

5.2. Convex Relaxations

So far we focused on theoretical ERM aspects only. Doing so, we made no assumptions about the computability of the step $h_i = \mathrm{argmin}_{h' \in \mathcal{C}} f_{h_{i-1}}(h')$ in Corollary 4 (Step 3 in Algoroithm 1). Although ERM results are interesting in their own right, we take an additional step and consider convex relaxations.

Instead of optimizing $\mathrm{er}_{\mathcal{D}}(h)$ over the set $\mathcal{C}$, assume we are interested in optimizing $\widetilde{\mathrm{er}}_{\mathcal{D}}(\tilde{h})$ over $\tilde{h} \in \tilde{\mathcal{C}}$, where $\tilde{\mathcal{C}}$ is a convex set of functions from $\mathcal{X}$ to $\mathbf{R}$. Also assume there is a mapping $\phi : \tilde{\mathcal{C}} \mapsto \mathcal{C}$ which is used as a "rounding" procedure. For example, in the setting of Section 5.1 the set $\tilde{\mathcal{C}}$ consists of all vectors $\mathbf{w} \in \mathbf{R}^d$, and the rounding method $\phi : \tilde{\mathcal{C}} \mapsto \mathcal{C}$ converts $\mathbf{w}$ to a permutation π satisfying (5.1). When optimizing in $\tilde{C}$, one conveniently works with a convex relaxation $\widetilde{\mathrm{er}}_{\mathcal{D}} : \tilde{C} \to \mathbf{R}^+$ as surrogate for the discrete loss $\mathrm{er}_{\mathcal{D}}$, defined as follows

$$\widetilde{\mathrm{er}}_{\mathcal{D}}(\tilde{h}) = \mathbf{E}_{(X,Y)\sim\mathcal{D}}\left[\tilde{L}\left(\tilde{h}(X), Y\right)\right]. \tag{5.3}$$

where $\tilde{L} : \mathbf{R} \times \{0,1\} \mapsto \mathbf{R}^+$ is some function convex in the first argument, and satisfying

$$\mathbf{1}_{(\phi(\tilde{h}))(X)\neq Y} \leq c\tilde{L}\left(\tilde{h}(X), Y\right)$$

for all $\tilde{h} \in \mathcal{C}$ and $X \in \mathcal{X}$, where $c > 0$ is some constant. In words, this means that $\tilde{L}$ upper bounds the discrete loss (up to a factor of c). A typical choice for the example in Section 5.1 would be to define for all $\mathbf{w} \in \tilde{C}$ and $X = (u, v) \in \mathcal{X}$: $\mathbf{w}(X) = \langle \mathbf{w}, \mathbf{u} - \mathbf{v} \rangle$, and $\tilde{L}(a, b) = \max\{1 - a(2b - 1), 0\}$. Using this choice, (5.3) becomes the famous SVMRank with the hinge loss relaxation.

We now have a natural extension of relative regret: $\widetilde{\mathrm{reg}}_{\tilde{h}}(\tilde{h}') = \widetilde{\mathrm{er}}_{\mathcal{D}}(\tilde{h}') - \widetilde{\mathrm{er}}_{\mathcal{D}}(\tilde{h})$. By our assumptions on convexity, $\widetilde{\mathrm{reg}}_{\tilde{h}} : \tilde{C} \mapsto \mathbf{R}^+$ can be efficiently optimized. We now say that $f : \tilde{C} \mapsto \mathbf{R}^+$ is an (ε, μ)-SRRA with respect to $\tilde{h} \in \tilde{C}$ if for all $\tilde{h}' \in \tilde{C}$,

$$\left| \mathrm{reg}_{\tilde{h}'}(\tilde{h}') - f(\tilde{h}') \right| \leq \varepsilon \left(\mathrm{dist}\big(\phi(\tilde{h}), \phi(\tilde{h}')\big) + \mu \right) .$$

If $\mu = 0$ then we simply say that f is an ε-SRRA. The following is an analogue to Corollary 4:

Theorem 14 *Let $\tilde{h}_0, \tilde{h}_1, \tilde{h}_2, \ldots$ be a sequence of hypotheses in $\tilde{\mathcal{C}}$ such that for all $i \geq 1$, $\tilde{h}_i = \mathrm{argmin}_{\tilde{h}' \in \tilde{\mathcal{C}}} f_{i-1}(\tilde{h}')$, where f_{i-1} is an (ε, μ)-SRRA with respect to $\tilde{h}_{i-1}$. Then for all $i \geq 1$,*

$$\widetilde{\mathrm{er}}_{\mathcal{D}}(h_i) = (1 + O(\varepsilon))\, \tilde{\nu} + O(\varepsilon^i) \widetilde{\mathrm{er}}_{\mathcal{D}}(h_0) + O(\varepsilon\mu) ,$$

where $\tilde{\nu} = \inf_{\tilde{h} \in \tilde{C}} \widetilde{\mathrm{er}}_{\mathcal{D}}(\tilde{h})$ and the O-notations may hide constants that depend on c.

The proof is very similar to that of Corollary 4, and we omit the details. It can be shown that the sampling techniques used for constructing an ε-SRRA from Section 4.1.1 can be used for constructing an ε-SRRA for the SVMRank relaxed version as well, as long as $\mathcal{C}$ is restricted to bounded vectors $\mathbf{w}$ and all the feature vectors $\mathbf{v}$ corresponding to $v \in V$ are bounded as well. The conclusion is that we can solve SVMRank, in polynomial time, to within an error of $(1 + \varepsilon)\tilde{\nu}$ using only $O\big(n \,\mathrm{poly}(\log n, \varepsilon^{-1})\big)$ preference queries. We leave the details of this simple extension to the full version.

6. Conclusions and Future Work

In this work we showed that being able to estimate the relative regret function using carefully biased sampling methods can yield query efficient active learning algorithms. We showed that such estimations can be obtained when the only assumptions we make are bounds on the disagreement coefficient and the VC dimension. This leads to active learning algorithms that almost match the best known using the same assumptions. On the other hand, we presented two problems of vast interest (mostly outside but increasingly inside the active learning community), for which a direct analysis of the relative regret function produced better active learning strategies. The two problems we studied are concerned with learning relations over a ground set, where one problem dealt order relations and the other with equivalence relations (with bounded number of equivalence classes). In both problems our query complexity bounds had an undesirable factors of ε^{-3} which we believe should be reduced to ε^{-2} using more advanced measure concentration tools. We leave this to future work. It would also be interesting to identify other problems for which our approach yields active learning algorithms with faster than previously known convergence rates. Immediate candidates are hierarchical clustering and metric learning. Finally, for LRPP, we discussed a practical scenario in which the ground set is endowed with feature vectors. We showed how to take the underlying geometry into account in our framework. We did not do so for clustering with side information. The

work of Eriksson et al. (2011) indicates that incorporating geometric information into our analysis is a fruitful direction to pursue.

Our work made no assumptions on the noise, except maybe for its magnitude. Another promising future research direction would be to incorporate various standard noise assumptions known to improve active learning rates (especially the model of Mammen and Tsybakov, 1999; Tsybakov, 2004) within our setting.

Acknowledgments

We thank Alekh Agarwal, Miroslav Dudik, Ran El-Yaniv, Sariel Har-Peled, John Langford, Rob Schapire, Masashi Sugiyama, and Yair Weiner for helpful discussions.

References

N. Ailon. An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *Journal of Machine Learning Research*, 13:137–164, 2012.

N. Ailon, B. Chazelle, S. Comandur, and D. Liu. Estimating the distance to a monotone function. *Random Struct. Algorithms*, 31(3):371–383, 2007.

N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: Ranking and clustering. *Journal of the ACM*, 55(5):23:1–23:27, Oct. 2008.

N. Ailon, R. Begleiter, and E. Ezra. Active learning using smooth relative regret approximations with applications (full version). In *arXiv:1110.2136*, 2012.

N. Alon. Ranking tournaments. *SIAM Journal on Discrete Mathematics*, 20, 2006.

F. R. Bach. Active learning for misspecified generalized linear models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 65–72. MIT Press, Cambridge, MA, 2007.

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *ICML*, pages 65–72, 2006.

M.-F. Balcan, A. Z. Broder, and T. Zhang. Margin based active learning. In *COLT*, pages 35–50, 2007.

M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *COLT*, pages 45–56, 2008.

N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56:89–113, 2004.

S. Basu. *Semi-supervised Clustering: Probabilistic Models, Algorithms and Experiments*. PhD thesis, Department of Computer Sciences, University of Texas at Austin, 2005.

A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *ICML*, 2009.

A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *NIPS*, 2010.

M. Braverman and E. Mossel. Noisy sorting without resampling. In *SODA*, pages 268–276, 2008.

R. Castro, R. Willett, and R. Nowak. Faster rates in regression via active learning. In *NIPS*, 2005.

R. M. Castro and R. D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear classification and selective sampling under low noise conditions. In *NIPS*, pages 249–256, 2008.

G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Learning noisy linear classifiers via adaptive and selective sampling. *Machine Learning*, 83(1):71–102, 2011.

N. Cesa-Bianchi, C. Gentile, F. Vitale, and G. Zappella. Active learning on trees and graphs. In *COLT*, pages 320–332, 2010.

M. Charikar and A. Wirth. Maximizing quadratic programs: Extending grothendieck's inequality. In *FOCS*, pages 54–60. IEEE Computer Society, 2004.

D. Cohn, R. Caruana, and A. Mccallum. Semi-supervised clustering with user feedback. unpublished manuscript, 2000. URL http://www.cs.umass.edu/~mccallum/papers/semisup-aaai2000s.ps.

D. Coppersmith, L. K. Fleischer, and A. Rurda. Ordering by weighted number of wins gives a good ranking for weighted tournaments. *ACM Trans. Algorithms*, 6:55:1–55:13, July 2010.

S. Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.

S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *ICML*, pages 208–215, 2008.

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *NIPS*, 2007.

M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars. *Computational geometry: Algorithms and applications*. Springer-Verlag, Berlin, 3rd edition, 2008.

A. Demiriz, K. Bennett, and M. J. Embrechts. Semi-supervised clustering using genetic algorithms. In *In Artificial Neural Networks in Engineering (ANNIE-99*, pages 809–814. ASME Press, 1999.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

B. Eriksson, G. Dasarathy, A. Singh, and R. D. Nowak. Active clustering: Robust and efficient hierarchical clustering using adaptively selected similarities. *Journal of Machine Learning Research - Proceedings Track*, 15:260–268, 2011.

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

I. Giotis and V. Guruswami. Correlation clustering with a fixed number of clusters. *Theory of Computing*, 2(1):249–266, 2006.

S. Halevy and E. Kushilevitz. Distribution-free property-testing. *SIAM J. Comput.*, 37(4):1107–1138, 2007.

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

S. Hanneke. Adaptive rates of convergence in active learning. In *COLT*, 2009.

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

S. Hanneke and L. Yang. Negative results for active learning with convex losses. *Journal of Machine Learning Research - Proceedings Track*, 9:321–325, 2010.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Control*, 100(1):78–150, Sept. 1992.

K. G. Jamieson and R. Nowak. Active ranking using pairwise comparisons. In *NIPS 24*, pages 2240–2248, 2011.

M. Kääriäinen. Active learning in the non-realizable case. In *ALT*, pages 63–77, 2006.

C. Kenyon-Mathieu and W. Schudy. How to rank with few errors. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, STOC '07, pages 95–103, 2007.

D. Klein, S. D. Kamvar, and C. D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *ICML*, pages 307–314, 2002.

V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.

Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62:2001, 2000.

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27:1808–1829, 1999.

S. Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13:67–90, 2012.

F. Orabona and N. Cesa-Bianchi. Better algorithms for selective sampling. In *ICML*, pages 433–440, 2011.

K. Radinsky and N. Ailon. Ranking from pairs and triplets: information quality, evaluation methods and query complexity. In *WSDM*, pages 105–114, 2011.

B. Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

O. Shamir and N. Tishby. Spectral clustering on a budget. *Journal of Machine Learning Research - Proceedings Track*, 15:661–669, 2011.

R. Shamir, R. Sharan, and D. Tsur. Cluster graph modification problems. *Discrete Applied Math*, 144:173–182, nov 2004.

M. Sugiyama. Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, 7:141–166, 2006.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32: 135–166, 2004.

K. Voevodski, M.-F. Balcan, H. Röglin, S.-H. Teng, and Y. Xia. Active clustering of biological sequences. *Journal of Machine Learning Research*, 13:203–225, 2012.

L. Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12:2269–2292, 2011.

E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, 2002.

L. Yang, S. Hanneke, and J. G. Carbonell. Bayesian active learning using arbitrary binary valued queries. In *ALT*, pages 50–58, 2010.

L. Yang, S. Hanneke, and J. G. Carbonell. The sample complexity of self-verifying bayesian active learning. *Journal of Machine Learning Research - Proceedings Track*, 15:816–822, 2011.

Appendix A. Proof of Theorem 3

Let $h^* \triangleq \operatorname{argmin}_{h' \in \mathcal{C}} \operatorname{reg}_h(h')$. Applying the definition of ε-SRRA we have:

$$\begin{aligned}
\operatorname{er}_{\mathcal{D}}(h_1) \leq & \operatorname{er}_{\mathcal{D}}(h) + f(h_1) + \varepsilon \operatorname{dist}(h, h_1) + \varepsilon\mu \\
\leq & \operatorname{er}_{\mathcal{D}}(h) + f(h^*) + \varepsilon \operatorname{dist}(h, h_1) + \varepsilon\mu \\
\leq & \operatorname{er}_{\mathcal{D}}(h) + \nu - \operatorname{er}_{\mathcal{D}}(h) + \varepsilon \operatorname{dist}(h, h^*) + \varepsilon \operatorname{dist}(h, h_1) + 2\varepsilon\mu \\
\leq & \nu + \varepsilon\Big(2\operatorname{dist}(h, h^*) + \operatorname{dist}(h_1, h^*)\Big) + 2\varepsilon\mu. \qquad \text{(A.1)}
\end{aligned}$$

The first inequality is from the definition of (ε, μ)-SRRA, the second is from the fact that h_1 minimizes $f(\cdot)$ by construction, the third is again from definition of (ε, μ)-SRRA and from the definition of the relative regret function reg_h, the fourth is by the triangle inequality. Now, clearly for any two hypotheses $g, g' \in C$ we have that $\operatorname{dist}(g, g') \leq \operatorname{er}_{\mathcal{D}}(g) + \operatorname{er}_{\mathcal{D}}(g')$ by the triangle inequality. The proof is completed by plugging $\operatorname{dist}(h, h^*) \leq \operatorname{er}_{\mathcal{D}}(h) + \nu$, and $\operatorname{dist}(h_1, h^*) \leq \operatorname{er}_{\mathcal{D}}(h_1) + \nu$ into Equation A.1, subtracting $\varepsilon \cdot \operatorname{er}_{\mathcal{D}}(h_1)$ from both sides, and dividing by $(1 - \varepsilon)$.

Appendix B. Proof of Corollary 4

Applying Theorem 3 with h_i and h_{i-1}, we have $\mathrm{er}_{\mathcal{D}}(h_i) = (1+O(\varepsilon))\,\nu + O\left(\varepsilon \cdot \mathrm{er}_{\mathcal{D}}(h_{i-1})\right) + O(\varepsilon\mu)$. Solving this recursion, one gets $\mathrm{er}_{\mathcal{D}}(h_i) = \sum_{j=1}^{i} \varepsilon^{j-1}\,(1+O(\varepsilon))\,\nu + O(\varepsilon^i)\cdot \mathrm{er}_{\mathcal{D}}(h_0) + O\left(\sum_{j=1}^{i}\varepsilon^j\right)\mu$. The result follows easily by bounding geometric sums.

Appendix C. Proof of Theorem 5

Proof Let $h \in \mathcal{X} \mapsto \mathbf{R}^+$ be some function, and let $\mu_h = E_{X\sim\mathcal{D}_{\mathcal{X}}}[h(X)]$. Let $X_1, \ldots, X_m$ denote i.i.d. draws from $\mathcal{D}_{\mathcal{X}}$, and let $\hat{\mu}_h \triangleq \frac{1}{m}\cdot\sum_{i=1}^{m} h(X_i)$ denote the empirical average. Let $\kappa > 0$ be an adjustable parameter. We are going to use the following measure of distance between μ_h and its estimator $\hat{\mu}_h$, to determine how far the latter is from the true expectations:

$$d_\kappa(\mu_h, \hat{\mu}_h) = \frac{|\mu_h - \hat{\mu}_h|}{\mu_h + \hat{\mu}_h + \kappa}.$$

This measure corresponds to a relative error when approximating μ by $\hat{\mu}_h$ (called *relative ε-approximations* in Haussler, 1992). Indeed, let $\varepsilon > 0$ be our approximation ratio, and put $d_\kappa(\mu_h, \hat{\mu}_h) < \varepsilon$. This easily yields

$$|\mu_h - \hat{\mu}_h| < \frac{2\varepsilon}{1-\varepsilon}\cdot\mu_h + \frac{\varepsilon}{1-\varepsilon}\cdot\kappa. \tag{C.1}$$

In other words, this implies that $|\mu_h - \hat{\mu}_h| < O(\varepsilon)(\mu_h + \kappa)$.

Let us fix a parameter $0 < \delta < 1$. Assume $\mathcal{C}$ is a set of $\{0,1\}$ valued functions on $\mathcal{X}$ of VC dimension d. Li et al. (2000) have shown that if one samples $m \triangleq c(\varepsilon^{-2}\kappa^{-1}(d\log\kappa^{-1} + \log\delta^{-1}))$ examples as above, then (C.1) holds uniformly for all $h \in \mathcal{C}$ with probability at least $1-\delta$.

For any h', define

$$\begin{aligned}
R^{++}_{h'} &= \{X \in \mathcal{X} :\ h'(X) = Y(X) = 1 \text{ and } h(X) = 0\}\\
R^{+-}_{h'} &= \{X \in \mathcal{X} :\ h'(X) = 1 \text{ and } h(X) = Y(X) = 0\}\\
R^{-+}_{h'} &= \{X \in \mathcal{X} :\ h'(X) = 0 \text{ and } h(X) = Y(X) = 1\}\\
R^{--}_{h'} &= \{X \in \mathcal{X} :\ h'(X) = Y(X) = 0 \text{ and } h(X) = 1\}\ .
\end{aligned}$$

Observe that the set $\{X \in \mathcal{X} : h(X) \neq h'(X)\}$ equals to the disjoint union of $R^{++}_{h'}$, $R^{+-}_{h'}$, $R^{-+}_{h'}$ and $R^{--}_{h'}$. For each $i = 0, \ldots, L$ and $b \in \{++, +-, -+, --\}$ let $R^b_{h',i} = R^b_{h'} \cap \mathcal{X}_i$. Let $\mathcal{R}^b_i = \{R^b_{h',i} : h' \in \mathcal{C}\}$. It is easy to verify that the VC dimension of the range spaces $\left(\mathcal{X}_i, \mathcal{R}^b_i\right)$ is at most d. Each set in $\mathbf{R}^b_i$ is an intersection of a set in $\mathcal{C}^*$ with some fixed set.

For any $R \subseteq \mathcal{X}_i$ let $\rho_i(R) = \Pr_{X\sim\mathcal{D}_{\mathcal{X}}|\mathcal{X}_i}[X \in R]$, and $\hat{\rho}_i(R) = m^{-1}\sum_{j=1}^{m}\mathbf{1}_{X_{i,j}\in R}$. Note that $\hat{\rho}_i(R)$ is an unbiased estimator of $\rho_i(R)$.

By the choice of m, inequality (C.1), and the properties discussed earlier Section 3.1 we have that with probability at least $1-\delta/L$, for all $R \in \mathcal{R}^{++}_i \cup \mathcal{R}^{+-}_i \cup \mathcal{R}^{-+}_i \cup \mathcal{R}^{--}_i$,

$$|\rho_i(R) - \hat{\rho}_i(R)| = O(\varepsilon)\cdot\left(\rho_i(R) + \theta^{-1}\right), \tag{C.2}$$

and by the probability union bound we obtain that this uniformly holds for all $i = 0, \dots L$, with probability at least $1 - \delta$.

Now fix $h' \in \mathcal{C}$ and let $r = \mathrm{dist}(h, h')$. Let $r(i) = \lceil \log(r/\mu) \rceil$. By the definition of $\mathcal{X}_i$, $h(X) = h'(X)$ for all $X \in \mathcal{X}_i$ whenever $i > r(i)$. We can therefore decompose $\mathrm{reg}_h(h')$ as:

$$\begin{aligned}
\mathrm{reg}_h(h') &= \mathrm{er}_{\mathcal{D}}(h') - \mathrm{er}_{\mathcal{D}}(h) \\
&= \sum_{i=0}^{L} \eta_i \cdot \left(\Pr_{X \sim \mathcal{D}_{\mathcal{X}} | \mathcal{X}_i} [Y(X) \neq h'(X)] - \Pr_{X \sim \mathcal{D}_{\mathcal{X}} | \mathcal{X}_i} [Y(X) \neq h(X)] \right) \\
&= \sum_{i=0}^{i(r)} \eta_i \cdot \left(\Pr_{X \sim \mathcal{D}_{\mathcal{X}} | \mathcal{X}_i} [Y(X) \neq h'(X)] - \Pr_{X \sim \mathcal{D}_{\mathcal{X}} | \mathcal{X}_i} [Y(X) \neq h(X)] \right) \\
&= \sum_{i=0}^{i(r)} \eta_i \cdot \left(- \rho_i(R_{h'}^{++}) + \rho_i(R_{h'}^{+-}) + \rho_i(R_{h'}^{-+}) - \rho_i(R_{h'}^{--}) \right) .
\end{aligned}$$

On the other hand, we similarly have that

$$f(h') = \sum_{i=0}^{i(r)} \eta_i \cdot \left(- \hat{\rho}_i(R_{h'}^{++}) + \hat{\rho}_i(R_{h'}^{+-}) + \hat{\rho}_i(R_{h'}^{-+}) - \hat{\rho}_i(R_{h'}^{--}) \right) .$$

Combining, we conclude using (C.2) that

$$\left| \mathrm{reg}_h(h') - f(h') \right| \leq O \left(\varepsilon \sum_{i=0}^{i(r)} \eta_i \cdot \left(\rho_i(R_{h'}^{++}) + \rho_i(R_{h'}^{+-}) + \rho_i(R_{h'}^{-+}) + \rho_i(R_{h'}^{--}) + 4\theta^{-1} \right) \right) \quad \text{(C.3)}$$

But now notice that $\sum_{i=0}^{i(r)} \eta_i \cdot \left(\rho_i(R_{h'}^{++}) + \rho_i(R_{h'}^{+-}) + \rho_i(R_{h'}^{-+}) + \rho_i(R_{h'}^{--}) \right)$ equals r, since it corresponds to those elements $X \in \mathcal{X}$ on which h, h' disagree. Also note that $\sum_{i=0}^{r(i)} \eta_i$ is at most $2 \max \left\{ \Pr_{\mathcal{D}_{\mathcal{X}}} \left[\mathrm{DIS} \left(\mathcal{B}(h, r) \right) \right], \Pr_{\mathcal{D}_{\mathcal{X}}} \left[\mathrm{DIS} \left(\mathcal{B}(h, \mu) \right) \right] \right\}$. By the definition of θ, this implies that the RHS of (C.3) is bounded by $\varepsilon(r + \mu)$, as required by the definition of (ε, μ)-SRRA.[11] ■

Appendix D. Notes

"Learning to Rank" takes various forms in theory and practice of learning, as well as in combinatorial optimization. In all versions, the goal is to order a set V based on constraints.

A large body of learning literature considers the following scenario: For each $v \in V$ there is a label on some discrete ordinal scale (say, $\{1, 2, 3, 4, 5\}$, as in hotel/restaurant star quality), and the goal is to learn how to order V so as to respect induced pairwise preferences. That is to say, if u had a label of 5 ("very good") and v had a label of 1 ("very bad"), then any ordering that places v ahead of u is penalized. Note that even if the labels are noisy, the induced pairwise preferences

11. The O-notation disappeared because we assume that the constants are properly chosen in the definition of the sample size m.

here are always transitive, hence no combinatorial problem arises. Our work does not deal with this problem.

When the basic unit of information consists of pairwise preferences over pairs $u, v \in V$, then the problem becomes combinatorially interesting. In case all quadratically many pairwise preferences are given for free, the corresponding optimization problem is known as *Minimum Feedback Arc-Set in Tournaments* (MFAST). (A maximization version exists as well.) MFAST is NP-hard (Alon, 2006). Recently Kenyon-Mathieu and Schudy (2007) show a (non query efficient) PTAS for this problem. Several important recent works address the challenge of approximating the minimum feedback arc-set problem (Ailon et al., 2008; Braverman and Mossel, 2008; Coppersmith et al., 2010).

Here we consider a query efficient variant of the problem, in which each preference comes at a cost, and the goal is to produce a competitive solution while reducing the preference-query overhead. Other very recent work consider similar settings (Jamieson and Nowak, 2011; Ailon, 2012). Our main result Corollary 8 is a slight improvement over the main result in (Ailon, 2012) in query complexity, but it provides another significant improvement. The querying strategy of Ailon (2012) is based on a divide and conquer strategy. The weakness of such a strategy can be explained by considering an example in which we want to search a restricted set of permutations (e.g., the setting of Section 5.1). When dividing and conquering, the algorithm in (Ailon, 2012) is doomed to search a cartesian product of two permutations spaces (left and right). There is no guarantee that there even exists a permutation in the restricted space that respects this division. In our querying algorithm this limitation is lifted.

Clustering with side information is a fairly new variant of clustering first described, independently, by Demiriz et al. (1999), and Ben-Dor et al. (1999). In the machine learning community it is also widely known as *semi-supervised clustering*. There are a few alternatives for the form of feedback providing the side-information. The most natural ones are the single item labels (e.g., Demiriz et al., 1999), and the pairwise constraints (e.g., Ben-Dor et al., 1999).

Here we consider pairwise side information: "must"/"cannot" link for pairs of elements $u, v \in V$. Each such information bit comes at a cost, and must be treated frugally. In a combinatorial optimization theoretical setting known as *correlation clustering* there is no input cost overhead, and similarity information for all (quadratically many) pairs is available. The goal there is to optimally clean the noise (nontransitivity). Correlation clustering was defined in (Bansal et al., 2004), and also in (Shamir et al., 2004) under the name *cluster editing*. Constant factor approximations are known for various minimization versions of this problems (Charikar and Wirth, 2004; Ailon et al., 2008). A PTAS is known for a minimization version in which the number of clusters is fixed to be k (Giotis and Guruswami, 2006), as in our setting.

In machine learning, there are two main approaches for utilizing pairwise side information. In the first approach, this information is used to fine tune or learn a *distance* function, which is then passed on to any standard clustering algorithm such as k-means or k-medians (see, e.g., Klein et al., 2002; Xing et al., 2002; Cohn et al., 2000; Shamir and Tishby, 2011; Voevodski et al., 2012). The second approach, which is more related to our work, modifies the clustering algorithms's objective so as to incorporate the pairwise constraints (see, e.g., Basu, 2005; Eriksson et al., 2011). Basu (2005) in his thesis, which also serves as a comprehensive survey, has championed this approach in conjunction with k-means, and hidden Markov random field clustering algorithms. In our work we isolate the use of information coming from pairwise clustering constraints, and separate it from the geometry of the problem. In future work it would be interesting to analyze our framework in

conjunction with the geometric structure of the input. Interestingly Eriksson et al. (2011) study active learning for clustering using the geometric input structure. Unlike our setting, they assume either no noise or Bayesian noise.