Journal of Machine Learning Research 15 (2014) 885-920 Submitted 9/12; Revised 6/13; Published 3/14

Active Learning Using Smooth Relative Regret Approximations with Applications

Nir Ailon NAILON@CS.TECHNION.AC.IL
Ron Begleiter RONBEG@CS.TECHNION.AC.IL
Department of Computer Science
Taub Building
Technion Israel Institute of Technology
Haifa 32000, Israel

Esther Ezra ESTHER@CIMS.NYU.EDU
Coutrant Institute of Mathematical Science
New York University
251 Mercer Street
New York, NY, 10012 USA

Editor: Yoav Freund

Abstract

The disagreement coefficient of Hanneke 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 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 applications of interest are: Learning to rank from pairwise preferences, and clustering with side information (a.k.a. semi-supervised clustering).

The tool we use is based on the learner's ability to compute an estimator of the difference between the loss of any hypothesis and some fixed "pivotal" hypothesis to within an absolute error of at most ε times the disagreement measure (ℓ_1 distance) between the two hypotheses. We prove that such an estimator implies the existence of a learning algorithm which, at each iteration, reduces its in-class 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.

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

1. Introduction

An *active learner* selects the instances from which it learns, contrary to standard *PAC learning*. In the streaming setting, active learners 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 labels are requested. 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; 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 the survey by Settles (2009) for a further discussion about 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 allow 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 only bounded VC dimension d and disagreement coefficient θ can roughly be stated as follows: If the sought (in-class) excess risk μ has the same order of magnitude as the optimal error ν or larger, then the number of required queries is roughly $\widetilde{O}(\theta d \log(1/\mu))$.[1] Otherwise, the number is roughly $\widetilde{O}(\theta d \nu^2/\mu^2)$. Note that this results makes no assumption on the noise (except maybe for its magnitude). Better results can be achieved 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 the optimum, 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 around the optimum is not unanimous on. This means that for minimizing regret, one may restrict attention to hypotheses laying in iteratively shrinking *version spaces* and to instances in the corresponding disagreement regions, which are shrinking in tandem with the version spaces if the disagreement coefficient is small. As pointed out in Beygelzimer et al. (2010), ignoring hypotheses outside the version space is brittle business, because an error in computation of the version space results in a failure of the algorithm. 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 of hypotheses ("relative regrets") is used vastly in numerous theoretical work on active learning, but not attached directly. In this work we argue that a careful construction of empirical processes uniformly estimating the relative regret of all hypotheses with respect to a fixed "pivotal" hypothesis yields fast active learning rates. We call such constructions "SRRA" (Smooth Relative Regret Approximations).

1. The $\widetilde{O}$ 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 also show that 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. Nevertheless, our SRRA-based iterative method does *not* need to compute or restrict itself to shrinking version spaces. This is supported by presenting two fundamental pool based learning problems for which *direct* SRRA constructions yield superior active learning rates, whereas approaches that exploit the disagreement coefficient only (in the sense presented in Section 3), 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.

We note that another important line of design and analysis of active learning algorithms makes certain structural or Bayesian 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 low disagreement coefficients imply efficient SRRAs. Then, we present our two main applications of SRRA that lie beyond the scope of disagreement coefficients: (i) learning to rank from pairwise preferences (LRPP) (Section 4), and (ii) clustering with side information in (Section 5). In Section 6 we present additional results and practical considerations, and, in particular, the application of our method with convex relaxation in case the corresponding ERM problems are too difficult (computationally) to optimally solve. We conclude in Section 7 and suggest future directions.

2. Definitions, Notations, and Core Results

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 (only) that each label Y is a deterministic function of X, so that if $X \sim \mathcal{D}_{\mathcal{X}}$ 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

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

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 $h_1, h_2 \in \mathcal{C}$ 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

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

2.1 The Disagreement Coefficient

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

$$\theta_h = \sup_{r>0} \frac{\Pr_{\mathcal{D}_{\mathcal{X}}}\left[\mathrm{DIS}\left(\mathcal{B}(h,r)\right)\right]}{r} \ , \tag{1}$$

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

$$\theta = \sup_{h\in\mathcal{C}} \sup_{r>0} \frac{\Pr_{\mathcal{D}_{\mathcal{X}}}\left[\mathrm{DIS}\left(\mathcal{B}(h,r)\right)\right]}{r} \ . \tag{2}$$

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 (1) and (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 $\mathrm{reg}_h : \mathcal{C} \mapsto \mathbb{R}$ 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 (in-class) excess risk function.

Definition 2 *Let $f : \mathcal{C} \mapsto \mathbb{R}$ 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 call f an ε-smooth relative regret approximation *with respect to h.*

Although the definition is general, we focus here on the pool based active learning setting. Intuitively, think of f as an empirical version of reg_h. The definition guides us to query labels such that we cover the whole spectrum of disagreement-instances w.r.t. the pivot h, while assuring corresponding estimation accuracies with granularity proportional to inverse distances from h. Intuitively, the condition supports holistic explore-exploit query strategies: We cover the whole range of error "types" (exploration), and at the same time put more querying efforts "near" the intermediate solution (exploitation). The following theorem and corollary constitute the main ingredient in our work. They show that a sequence of (ε,μ)-SRRA estimators define a competitive hypothesis.

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*

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

Proof Applying the definition of (ε, μ)-SRRA we have:

$$\begin{aligned}
\mathrm{er}_{\mathcal{D}}(h_1) \leq & \mathrm{er}_{\mathcal{D}}(h) + f(h_1) + \varepsilon \cdot \mathrm{dist}(h, h_1) + \varepsilon\mu \\
\leq & \mathrm{er}_{\mathcal{D}}(h) + f(h^*) + \varepsilon \cdot \mathrm{dist}(h, h_1) + \varepsilon\mu \\
\leq & \mathrm{er}_{\mathcal{D}}(h) + \nu - \mathrm{er}_{\mathcal{D}}(h) + \varepsilon \cdot \mathrm{dist}(h, h^*) + \varepsilon \cdot \mathrm{dist}(h, h_1) + 2\varepsilon\mu \\
\leq & \nu + \varepsilon\Big(2\mathrm{dist}(h, h^*) + \mathrm{dist}(h_1, h^*)\Big) + 2\varepsilon\mu \,. \qquad (3)
\end{aligned}$$

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 the definition of (ε, μ)-SRRA, and the definitions of h^* and reg_h, the fourth is by the triangle inequality. The proof is completed by plugging $\mathrm{dist}(h, h^*) \leq \mathrm{er}_{\mathcal{D}}(h) + \nu$, and $\mathrm{dist}(h_1, h^*) \leq \mathrm{er}_{\mathcal{D}}(h_1) + \nu$ into Equation 3, subtracting $\varepsilon \cdot \mathrm{er}_{\mathcal{D}}(h_1)$ from both sides, and dividing by $(1 - \varepsilon)$. ■

A simple inductive use of Theorem 3 proves the following corollary, bounding the excess risk of an ERM based active learning algorithm (see Algorithm 1 for corresponding pseudocode). The algorithm's query-complexity depends on the specific constructions of (ε, μ)-SRRA estimators. Note that this 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 $i \geq 0$,*

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

Proof 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. ■

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: **for** $i = 0, 1, \ldots, T-1$ **do**
2: $\quad h_{i+1} \leftarrow \mathrm{argmin}_{h' \in \mathcal{C}} f(h')$, where f is an (ε, μ)-smooth relative regret approximation with respect to h_i
3: **end for**
4: **return** h_T

We will show below problems of interest in which (ε, μ)-SRRA's with respect to a given hypothesis h can be obtained using labels at few randomly (and adaptively) selected points

X from the pool $\mathcal{X}$, 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 SRRAs

We show that a bounded uniform disagreement coefficient implies existence of query efficient (ε, μ)-SRRAs. This constitutes yet another proof of the usefulness of the disagreement coefficient in design of active learning algorithms, via Algorithm 1. Plugging the resulting query efficient (ε, μ)-SRRAs into our iterative SRRA method (Algorithm 1) provides an active learning algorithm with query complexity that matches the state-of-the-art, yet does not improve it. In the following sections, however, we design two other (ε, μ)-SRRAs constructions that yield active learning algorithms which beats the corresponding state-of-the-art query complexity guarantees.

3.1 The Construction

Returning to our problem, assume the uniform disagreement coefficient θ corresponding to $\mathcal{C}$ is finite and $\nu > 0$. Fix $h \in \mathcal{C}$ and let $L = \lceil \log \mu^{-1} \rceil$. Define $\mathcal{X}_0 = \mathrm{DIS}(\mathcal{B}(h, \mu))$ and for $i = 1, \ldots, L$ 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 $\mathcal{X}_i$, and δ a failure probability hyper-parameter. 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 $f : \mathcal{C} \mapsto \mathbb{R}$ of reg_h, as follows. For any $h' \in \mathcal{C}$ and $i = 0, 1, \ldots, L$ let

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

Our estimator is now defined as $f(h') = \sum_{i=0}^{L} f_i(h')$. The estimator $f(h')$ is an unbiased empirical counterpart of the relative regret in which each empirical error $\left(\mathbf{1}_{Y(X_{i,j}) \neq h'(X_{i,j})} - \mathbf{1}_{Y(X_{i,j}) \neq h(X_{i,j})}\right)$ is weighted with respect to η_i. The weights are depicted as different element sizes. Observe that the weight is inverse proportional to the probability of drawing $X_{i,j}$.

We next show that f is an (ε, μ)-SRRA with respect to h. This allows us to incorporate f into Algorithm 1 and gain a $(1+\varepsilon)\nu$ competitive hypothesis. Thus, the query complexity will boil down to the size of the sample defining the empirical relative-regret minimizer f.

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.*

Proof A main tool to be exploited in the proof is called *relative ε-approximations* due to Haussler (1992) and Li et al. (2000). It is defined as follows. Let $h \in \mathcal{X} \mapsto \mathbb{R}^+$ 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 $\hat{\mu}_h = \frac{1}{m} \cdot \sum_{i=1}^m h(X_i)$ denote the empirical average. Let $\kappa > 0$ be an adjustable parameter. We are going to use the following measure of distance between the true expectation μ_h and its estimator $\hat{\mu}_h$: $d_\kappa(\mu_h, \hat{\mu}_h) = \frac{|\mu_h - \hat{\mu}_h|}{\mu_h + \hat{\mu}_h + \kappa}$.

This measure corresponds to a relative error when approximating μ_h by $\hat{\mu}_h$. Indeed, let $\varepsilon > 0$ be our approximation ratio, and put $d_\kappa(\mu_h, \hat{\mu}_h) < \varepsilon$. This easily yields

$$|\mu_h - \hat{\mu}_h| < \frac{2\varepsilon}{1-\varepsilon} \cdot \mu_h + \frac{\varepsilon}{1-\varepsilon} \cdot \kappa\,. \tag{4}$$

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

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) show that if one samples

$$m = O\left(\varepsilon^{-2}\kappa^{-1}(d\log\kappa^{-1} + \log\delta^{-1})\right)$$

examples as above (with a sufficiently large constant of proportionality), then (4) holds uniformly for all $h \in \mathcal{C}$ with probability at least $1 - \delta$.

We consider the range space $(\mathcal{X}, \mathcal{C}^*)$, defined by

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

In 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[3] d, and fix $h \in \mathcal{C}$. Let $L = \lceil \log \mu^{-1} \rceil$, and recall the definition of the disagreement sets $\mathcal{X}_i$.

We now apply this definition of *relative ε-approximations*, and the corresponding results within our context. For any h', we define the following four sets of instances:

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

Observe that the set $\{X \in \mathcal{X} : h(X) \neq h'(X)\}$ equals 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 $\mathcal{R}_i^{\,b} = \{R^b_{h',i} : h' \in \mathcal{C}\}$. It is easy to verify that the VC dimension of the range spaces $\big(\mathcal{X}_i, \mathcal{R}_i^{\,b}\big)$ is at most d. Each set in $\mathcal{R}_i^{\,b}$ 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 $\rho_i(R)$.

3. The VC dimension of $(\mathcal{X}, \mathcal{C}^*)$ is the maximum cardinality of a subset $A \subseteq \mathcal{X}$ for which $\{A \cap r : r \in \mathcal{C}^*\}$ contains all subsets of A.

By the choice of m, inequality (4), and the assumptions on θ and ν we have that with probability at least $1-\delta/L$, for all $R \subseteq \mathcal{R}_i^{++} \cup \mathcal{R}_i^{+-} \cup \mathcal{R}_i^{-+} \cup \mathcal{R}_i^{--}$,

$$|\rho_i(R) - \hat{\rho}_i(R)| = O(\varepsilon) \cdot \left(\rho_i(R) + \theta^{-1}\right), \tag{5}$$

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

Now fix $h' \in \mathcal{C}$ and let $r = \mathrm{dist}(h, h')$. Let $i_r = \lceil \log(r/\mu) \rceil$. By the definition of $\mathcal{X}_i$, $h(X) = h'(X)$ for all $X \in \mathcal{X}_i$ whenever $i > i_r$. 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 (5) that

$$|\mathrm{reg}_h(h') - f(h')| \leq O\left(\varepsilon \sum_{i=0}^{i_r} \eta_i \cdot \left(\rho_i(R_{h'}^{++}) + \rho_i(R_{h'}^{+-}) + \rho_i(R_{h'}^{-+}) + \rho_i(R_{h'}^{--}) + 4\theta^{-1} \right) \right). \tag{6}$$

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 $\sum_{i=0}^{i_r} \eta_i$ 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 (6) is bounded by $\varepsilon(r + \mu)$, as required by the definition of (ε, μ)-SRRA.[4] ■

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

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

label queries.

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

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 order at least that of the optimal error ν, then we may take ε to be, say, $1/5$ and achieve the sought bound by constructing $(1/5, \nu)$-SRRA's using $O(\theta d(\log(1/\nu))(\log\theta))$, 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 is $O(\theta d(\log^2(1/\nu))(\log\theta))$, which is $O(\log(1/\nu))$ times the best known bounds using disagreement coefficient and VC dimension bounds only (e.g., Dasgupta et al., 2007; Beygelzimer et al., 2009).

Remark 7 (Using the uniform disagreement coefficient) *We 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 in both previously known results bounding query complexity using disagreement coefficient and VC dimension bounds as well as our result, the slight improvement described in Remark 1 applies. In other words, all arguments remain valid if we replace the supremums in (1) and (2) with* $\sup_{r\geq\nu}$.

Remark 8 (Computing the $\mathcal{X}_i$'s) *Note that we show how to compute $\mathcal{X}_i$* exactly *in polynomial time when dealing with linear hypotheses spaces. This is shown in Section 6.1.2 in the context of a ranking problem. Yet, it indicates that in certain cases $\mathcal{X}_i$ can be computed efficiently. Additionally note that the sets $\mathcal{X}_i$ are defined w.r.t. a pivot (i.e., given) hypothesis; thus, it is possible to estimate it, for example, using ideas similar to the nice works of Balcan et al. (2006) and Dasgupta et al. (2007).*

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

"Learning to Rank" takes various forms in theory and practice of learning, as well as in combinatorial optimization. In such problems, 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, and the goal is to learn how to order V so as to respect induced pairwise preferences. For example, a scale of $\{1, 2, 3, 4, 5\}$, as in hotel/restaurant star quality; where, if u has a label of 5 ("very good") and v has 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 here are always transitive, hence no combinatorial problem arises. Our work does not deal with this setting.

When the basic unit of information consists of 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).[5] MFAST has been shown to be NP-hard (Alon, 2006). Recently, Kenyon-Mathieu and Schudy (2007) showed a PTAS for this (passive

5. A maximization version of this problem exists as well.

learning) 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). Jamieson and Nowak (2011) consider a common scenario in which the alternatives can be characterized in terms of d real-valued features and the ranking obeys the structure of the Euclidean distances between such embeddings. They present an active learning algorithm that requires, using *average case analysis*, as few as $O(d \log n)$ labels in the noiseless case, and $O(d \log^2 n)$ labels under a certain *parametric* noise model. Our work uses worst-case analysis, and assumes an adversarial noise model. In this Section we analyze the pure combinatorial problem (not assuming any feature embeddings). In Section 6 we tackle the problem with linearly induced permutation over feature space embeddings.

Ailon (2012) consider the same setting as ours. Our main result Corollary 13 is a slight improvement over the main result of Ailon (2012) in query complexity, but it provides another significant improvement. Ailon (2012) uses a querying method that 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 6.1). 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.

4.1 Problem Definition

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 $u, v \in V$.[6] 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 it is *skew-symmetric*: $h(u,v) = 1 - h(v,u)$, and *transitive*: $h(u,z) \leq h(u,v) + h(v,z)$ for all distinct $u, v, z \in V$. This is 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 $N = n(n-1)$ is the number of all ordered pairs.

4.2 The Weakness of Using Disagreement Coefficient Arguments

We first demonstrate the weakness of a disagreement coefficient approach with respect to this problem. It has been shown in Ailon (2012) that the uniform disagreement coefficient

6. Note that we could have defined $\mathcal{X}$ to be unordered pairs of elements in V without making any assumption on Y. We chose this definition for convenience in what follows.

of $\mathcal{C}$ is $\Omega(n)$. 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 $n-1$ (e.g., Radinsky and Ailon, 2011). This is simply because there is always a labeling $Y(\cdot)$ over any set of $n = |V|$ *pairs* that defines preference cycles. Thus, the set of permutations $\mathcal{C}$ cannot shatter n pairs. 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 9 *A slight improvement can be obtained using the refined definition of disagreement coefficients of Remark 1. Observe that the uniform disagreement coefficient, as well as the disagreement coefficient at the optimal solution h^* becomes $\theta = \theta_{h^*} = O(1/\nu)$, if $\nu \geq \frac{1}{n}$.*[7] *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.*

Remark 10 *A natural question in this context is why the optimal hypothesis cannot be approximated by sampling preference-pairs uniformly at random. In other words, is it sufficient for our setting to apply a passive learning method? Do we really need to apply the more sophisticated active-learning machinery? Ailon (2012, Section 2) shows that applying plain Empirical Risk Minimization approach is doomed to query the entire pool. Moreover, he shows that even when the noise is zero the uniform sampling approach will w.h.p. have to query all $O(n^2)$ possible labels.*

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

4.3 Better SRRA for LRPP

Consider the following idea for creating an ε-SRRA for LRPP,[8] with respect to some fixed $\pi \in \mathcal{C}$. We start by defining the following sample size parameter:

$$p = O\left(\varepsilon^{-3} \log^3 n\right) \ . \tag{7}$$

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 $(2^i - 1)p < |\pi(u) - \pi(v)| < 2^{i+1}p$ where, abusing notation, $\pi(u)$ is the position of u in π. For example, $\pi(u)$ is 1 if u beats all other elements, and n if it is beaten by all others. From this set, choose a random sequence $R_{u,i} = (v_{u,i,1}, v_{u,i,2}, \ldots, v_{u,i,p})$ of p elements, each chosen uniformly and independently from $I_{u,i}$.[9]

7. Due to symmetry, the uniform disagreement coefficient here equals θ_h for any $h \in \mathcal{C}$.
8. Note that we can neglect the parameter μ because taking its value equal $1/n$ tantamount to zero.
9. A variant of this sampling scheme is as follows: For each pair (u, v), add it to the query-set with probability proportional to $\min\{1, p/|\pi(u) - \pi(v)|\}$. A similar scheme can be found in Ailon et al. (2007), Halevy and Kushilevitz (2007) and Ailon (2012) but the strong properties proven here were not known.

For distinct $u, v \in V$ and a permutation $\sigma \in \mathcal{C}$, let $\mathrm{cost}_{u,v}(\sigma)$ denote the contribution of the pair u, v to $\mathrm{er}_{\mathcal{D}}(\sigma)$, namely: $\mathrm{cost}_{u,v}(\sigma) = N^{-1}\mathbf{1}_{\sigma(u,v)\neq Y(u,v)}$. Let $\mathrm{reg}_{u,v|\sigma}$ denote the contribution of $\{u, v\} \in \mathcal{X}$ to $\mathrm{reg}_{\pi}(\sigma)$, that is

$$\mathrm{reg}_{u,v|\sigma} = 2\left(\mathrm{cost}_{u,v}(\sigma) - \mathrm{cost}_{u,v}(\pi)\right) . \tag{8}$$

Note the notation discards the dependency on π because it is assumed to be fixed. The use of factor 2 is because $\mathrm{cost}_{u,v} \equiv \mathrm{cost}_{v,u}$.

Our estimator $f(\sigma)$ of $\mathrm{reg}_{\pi}(\sigma) = \mathrm{er}_{\mathcal{D}}(\sigma) - \mathrm{er}_{\mathcal{D}}(\pi)$ is defined as

$$f(\sigma) = \frac{1}{2}\sum_{u\in V}\sum_{i=0}^{\lceil \log n\rceil}\frac{|I_{u,i}|}{p}\sum_{t=1}^{p}\mathrm{reg}_{u,v_{u,i,t}|\sigma} .$$

Clearly, $f(\sigma)$ is an unbiased estimator of $\mathrm{reg}_{\pi}(\sigma)$ for any σ. Our goal is to prove that $f(\sigma)$ is an ε-SRRA.

Theorem 11 *With probability at least $1 - n^{-3}$, the function f is an ε-SRRA with respect to π.*

Proof The main idea is to *decompose* the difference $|f(\sigma) - \mathrm{reg}_{\pi}(\sigma)|$ vis-a-vis corresponding pieces of $\mathrm{dist}(\sigma, \pi)$. The first half of the proof is devoted to definition of such distance "pieces." Then using counting and standard deviation-bound arguments we show that the decomposition is, with high probability, an ε-SRRA.

We start with a few definitions. Recall that for any $\pi \in \mathcal{C}$ and $u \in V$, $\pi(u)$ denotes the position of u in the unique permutation that π defines. For example, $\pi(u) = 1$ if u beats all other alternatives: $\pi(u, v) = 1$ for all $v \neq u$; Similarly $\pi(u) = n$ if u is beaten by all other alternatives. For any permutation $\sigma \in \mathcal{C}$, we define the corresponding *profile* of σ as the vector:[10]

$$\mathrm{prof}(\sigma) = \big(\sigma(u_1) - \pi(u_1), \sigma(u_2) - \pi(u_2), \ldots, \sigma(u_n) - \pi(u_n)\big).$$

Note that $\|\mathrm{prof}(\sigma)\|_1$ is $d_{\mathrm{SF}}(\sigma, \pi)$, the Spearman footrule distance between σ and π. For a subset V' of V, we let $\mathrm{prof}(\sigma)[V']$ denote the restriction of the vector $\mathrm{prof}(\sigma)$ to V'. Namely, the vector obtained by zeroing in $\mathrm{prof}(\sigma)$ all coordinates $v \notin V'$.

Now fix $\sigma \in \mathcal{C}$ and two distinct $u, v \in V$. Assume u, v is an inversion in σ with respect to π, and that $|\pi(u) - \pi(v)| = b$ for some integer b. Then either $|\pi(u) - \sigma(u)| \geq b/2$ or $|\pi(v) - \sigma(v)| \geq b/2$. We will "charge" the inversion to $\mathrm{argmax}_{z\in\{u,v\}}\left\{|\pi(z) - \sigma(z)|\right\}$.[11] For any $u \in V$, let $\mathrm{charge}_{\sigma}(u)$ denote the set of elements $v \in V$ such that (u, v) is an inversion in σ with respect to π, which is charged to u based on the above rule. The function $\mathrm{reg}_{\pi}(\sigma)$ can now be written as

$$\mathrm{reg}_{\pi}(\sigma) = \sum_{u\in V}\sum_{v\in\mathrm{charge}_{\sigma}(u)}\mathrm{reg}_{u,v|\sigma} ,$$

10. For the sake of definition assume an arbitrary indexing such that $V = \{u_i : i = 1, \ldots, n\}$.

11. Breaking ties using some canonical rule, for example, charge to the greater of u, v viewed as integers.

where $\mathrm{reg}_{u,v|\sigma}$ is defined in Equation 8. Indeed, any pair that is not inverted contributes nothing to the difference. Similarly, our estimator $f(\sigma)$ can be written as

$$f(\sigma) = \sum_{u \in U} \sum_{i=0}^{\lceil \log n \rceil} \frac{|I_{u,i}|}{p} \sum_{t=1}^{p} \mathrm{reg}_{u,v_{u,i,t}|\sigma} \cdot \mathbf{1}_{v_{u,i,t} \in \mathrm{charge}_\sigma(u)} \ .$$

Observe that we dropped the factor $1/2$ above because we count each pair $\{u, v\}$ only once.

For any even integer M let $U_{\sigma,M}$ denote the set of all elements $u \in V$ such that

$$M/2 < |\pi(u) - \sigma(u)| \leq M \ .$$

Let $U_{\sigma,\leq M}$ denote:

$$\bigcup_{M' \leq M} U_{\sigma,M'} \ .$$

From now on, we shall remove the subscript π, because it is held fixed. Consider the following restrictions of $\mathrm{reg}(\sigma)$ and $f(\sigma)$:

$$\mathrm{reg}(\sigma, M) = \sum_{u \in U_{\sigma,M}} \sum_{v \in \mathrm{charge}_\sigma(u)} \mathrm{reg}_{u,v|\sigma} \ , \tag{9}$$

$$f(\sigma, M) = \sum_{u \in U_{\sigma,M}} \sum_{i=0}^{\lceil \log n \rceil} \sum_{t=1}^{p} \frac{|I_{u,i}|}{p} \left(\mathrm{reg}_{u,v_{u,i,t}|\sigma} \cdot \mathbf{1}_{v_{u,i,t} \in \mathrm{charge}_\sigma(u)} \right) \ . \tag{10}$$

Clearly, $f(\sigma, M)$ is an unbiased estimator of $\mathrm{reg}(\sigma, M)$. Let $T_{\sigma,M}$ denote the set of all elements $u \in V$ such that $|\pi(u) - \sigma(u)| \leq \varepsilon M$. We further split the expressions in (9)-(10) as follows:

$$\mathrm{reg}(\sigma, M) = A(\sigma, M) + B(\sigma, M), \text{ and } f(\sigma, M) = \hat{A}(\sigma, M) + \hat{B}(\sigma, M),$$

where,

$$A(\sigma, M) = \sum_{u \in U_{\sigma,M}} \sum_{v \in \mathrm{charge}_\sigma(u) \cap \overline{T_{\sigma,M}}} \mathrm{reg}_{u,v|\sigma} \ ,$$

$$\hat{A}(\sigma, M) = \sum_{u \in U_{\sigma,M}} \sum_{i=0}^{\lceil \log n \rceil} \frac{|I_{u,i}|}{p} \sum_{t=1}^{p} \mathrm{reg}_{u,v_{u,i,t}|\sigma} \cdot \mathbf{1}_{v_{u,i,t} \in \mathrm{charge}_\sigma(u) \cap \overline{T_{\sigma,M}}} \ ,$$

$\overline{(\cdot)}$ is set complement in V, and $B(\sigma, M), \hat{B}(\sigma, M)$ are analogous with $T_{\sigma,M}$ instead of $\overline{T_{\sigma,M}}$, as follows:

$$B(\sigma, M) = \sum_{u \in U_{\sigma,M}} \sum_{v \in \mathrm{charge}_\sigma(u) \cap T_{\sigma,M}} \mathrm{reg}_{u,v|\sigma} \ ,$$

$$\hat{B}(\sigma, M) = \sum_{u \in U_{\sigma,M}} \sum_{i=0}^{\lceil \log n \rceil} \frac{|I_{u,i}|}{p} \sum_{t=1}^{p} \mathrm{reg}_{u,v_{u,i,t}|\sigma} \cdot \mathbf{1}_{v_{u,i,t} \in \mathrm{charge}_\sigma(u) \cap T_{\sigma,M}} \ .$$

We now estimate the deviation of $\hat{A}(\sigma, M)$ from $A(\sigma, M)$. Fix M. Notice that the expression $A(\sigma, M)$ is completely determined by non-zero elements of the vector $\mathrm{prof}(\sigma)[U_{\sigma,\leq M} \cap \overline{T_{\sigma,M}}]$. Let $J_{\sigma,M}$ denote the number of nonzeros in $\mathrm{prof}(\sigma)[U_{\sigma,M}]$. Each nonzero coordinate of $\mathrm{prof}(\sigma)[U_{\sigma,\leq M} \cap \overline{T_{\sigma,M}}]$ is bounded below by εM in absolute value by definition. Let $P(d, M)$ denote the number of possibilities for the vector $\mathrm{prof}(\sigma)[\overline{T_{\sigma,M}}]$ for σ running over all permutations satisfying $d_{\mathrm{SF}}(\sigma, \pi) = d$. We claim that

$$P(d, M) \leq n^{2d/(\varepsilon M)} \ .$$

Indeed, there can be at most $d/(\varepsilon M)$ nonzeros in $\mathrm{prof}(\sigma)[\overline{T_{\sigma,M}}]$, and each nonzero coordinate can trivially take at most n values. The bound follows.

Now fix integers d and J, and consider the subspace of permutations σ such that $J_{\sigma,M} = J$ and $d_{\mathrm{SF}}(\sigma, \pi) = d$. Define for each $u \in U_{\sigma,M}$, $i \in \lceil \log n \rceil$ and $t = 1, \ldots, p$ a random variable $X_{u,i,t}$ as follows

$$X_{u,i,t} = \frac{|I_{u,i}|}{p} \, \mathrm{reg}_{u, v_{u,i,t}|\sigma} \cdot \mathbf{1}_{v_{u,i,t} \in \mathrm{charge}_\sigma(u) \cap \overline{T_{\sigma,M}}} \ .$$

Clearly $\hat{A}(\sigma, M) = \sum_{u \in U_{\sigma,M}} X_{u,i,t}$. For any $u \in V$, let $i_u = \mathrm{argmax}_i \{|I_{u,i}| \leq 4M\}$, and observe that, by our charging scheme, $X_{u,i,t} = 0$ almost surely, for all $i > i_u$ and $t = 1 \ldots p$. Also observe that for all u, i, t, $|X_{u,i,t}| \leq 2N^{-1}|I_{u,i}|/p \leq 2^{i+1}/p$ almost surely. For a random variable X, we denote by $\|X\|_\infty$ the infimum over numbers α such that $X \leq \alpha$ almost surely. We conclude:

$$\sum_{u \in U_{\sigma,M}} \sum_{i=0}^{i_u} \sum_{t=1}^{p} \|X_{u,i,t}\|_\infty^2 \leq \sum_{u \in U_{\sigma,M}} \sum_{i=0}^{i_u} N^{-2} p 2^{2i+2}/p^2 \leq c_2 p^{-1} N^{-2} J M^2$$

for some global $c_2 > 0$. (We used a bound on the sum of a geometric series.) Using Hoeffding bound, we conclude that the probability that $\hat{A}(\sigma, M)$ deviates from its expected value of $A(\sigma, M)$ by more than some $s > 0$ is at most $\exp\{-s^2 p/(2c_2 J M^2 N^{-2})\}$. We conclude that the probability that $A(\sigma, M)$ deviates from its expected value by more than $\varepsilon d/(N \log n)$ is at most $\exp\{-c_1 \varepsilon^2 d^2 p/(J M^2 \log^2 n)\}$, for some global $c_1 > 0$. Hence, by taking $p = O(\varepsilon^{-3} d^{-1} M J \log^3 n)$, by union bounding over all $P(d, M)$ possibilities for $\mathrm{prof}(\sigma)[\overline{T_{\sigma,M}}]$, with probability at least $1 - n^{-7}$ simultaneously for all σ satisfying $J_{\sigma,M} = J$ and $d_{\mathrm{SF}}(\sigma, \pi) = d$,

$$|A(\sigma, M) - \hat{A}(\sigma, M)| \leq \varepsilon d/(N \log n) \ . \tag{11}$$

But note that, trivially, $JM \leq d$, hence our choice of p in (7) is satisfactory. Finally, union bound over the $O(n^3 \log n)$ possibilities for the values of J and d and $M = 1, 2, 4, ..$ to conclude that (11) holds for all permutations σ simultaneously, with probability at least $1 - n^{-3}$.

Consider now $\hat{B}(\sigma, M)$ and $B(\sigma, M)$. We will need to further decompose these two expressions as follows. For $u \in U_{\sigma,M}$, we define a disjoint cover $(T^1_{u,\sigma,M}, T^2_{u,\sigma,M})$ of $\mathrm{charge}_\sigma(u) \cap T_{\sigma,M}$ as follows. If $\pi(u) < \sigma(u)$, then

$$T^1_{u,\sigma,M} = \{v \in T_{\sigma,M} : \pi(u) + \varepsilon M < \pi(v) < \sigma(u) - \varepsilon M\} \ .$$

Otherwise,

$$T^1_{u,\sigma,M} = \{v \in T_{\sigma,M} : \sigma(u) + \varepsilon M < \pi(v) < \pi(u) - \varepsilon M\} \ .$$

Note that by definition, $T^1_{u,\sigma,M} \subseteq \mathrm{charge}_\sigma(u)$. The set $T^2_{u,\sigma,M}$ is thus taken to be

$$T^2_{u,\sigma,M} = (\mathrm{charge}_\sigma(u) \cap T_{\sigma,M}) \setminus T^1_{u,\sigma,M} \ .$$

The expressions $B(\sigma, M), \hat{B}(\sigma, M)$ now decompose as $B^1(\sigma, M)+B^2(\sigma, M)$ and $\hat{B}^1(\sigma, M)+\hat{B}^2(\sigma, M)$, respectively, as follows:

$$\begin{aligned}
B^1(\sigma,M) &= \sum_{u\in U_{\sigma,M}} \sum_{v\in T^1_{u,\sigma,M}} \mathrm{reg}_{u,v|\sigma} \ , \\
B^2(\sigma,M) &= \sum_{u\in U_{\sigma,M}} \sum_{v\in T^2_{u,\sigma,M}} \mathrm{reg}_{u,v|\sigma} \ , \\
\hat{B}^1(\sigma,M) &= \sum_{u\in U_{\sigma,M}} \sum_{i=0}^{\lceil \log n\rceil} \frac{|I_{u,i}|}{p} \sum_{t=1}^{p} \mathrm{reg}_{u,v_{u,i,t}|\sigma} \cdot \mathbf{1}_{v_{u,i,t}\in T^1(u,\sigma,M)} \ , \\
\hat{B}^2(\sigma,M) &= \sum_{u\in U_{\sigma,M}} \sum_{i=0}^{\lceil \log n\rceil} \frac{|I_{u,i}|}{p} \sum_{t=1}^{p} \mathrm{reg}_{u,v_{u,i,t}|\sigma} \cdot \mathbf{1}_{v_{u,i,t}\in T^2(u,\sigma,M)} \ .
\end{aligned}$$

Now notice that $B^1(\sigma, M)$ can be uniquely determined from $\mathrm{prof}(\sigma)[\overline{T_{\sigma,M}}]$. Indeed, in order to identify $T^1_{u,\sigma,M}$ for some $u \in U_{\sigma,M}$, it suffices to identify zeros in a subset of coordinates of $\mathrm{prof}(\sigma)[\overline{T_{\sigma,M}}]$, where the subset depends only on $\mathrm{prof}(\sigma)[\{u\}]$. Additionally, the value of $C_{u,v}(\sigma) - C_{u,v}(\pi)$ can be "read" from $\mathrm{prof}(\sigma)[\overline{T_{\sigma,M}}]$ (and, of course, $Y(u, v)$) if $v \in T^1_{u,\sigma,M}$. Hence, a Hoeffding bound and a union bound similar to the one used for bounding $|\hat{A}(\sigma, M) - A(\sigma, M)|$ can be used to bound (with high probability) the difference $|\hat{B}^1(\sigma, M) - B^1(\sigma, M)|$ uniformly for all σ and $M = 1, 2, 4, ...$, as well.

Bounding $|\hat{B}^2(\sigma, M) - B^2(\sigma, M)|$ can be done using the following simple claim.

Claim 12 *For $u \in V$ and an integer q, we say that the sampling is successful at (u, q) if the random variable*

$$\left|\left\{(i,t) : \pi(v_{u,i,t}) \in [\pi(u) + (1-\varepsilon)q,\ \pi(u) + (1+\varepsilon)q] \cup [\pi(u) - (1+\varepsilon)q,\ \pi(u) - (1-\varepsilon)q]\right\}\right|$$

is at most twice its expected value. We say that the sampling is successful if it is successful at all $u \in V$ and $q \le n$. If the sampling is successful, then uniformly for all σ and all $M = 1, 2, 4, ...$,

$$|\hat{B}^2(\sigma, M) - B^2(\sigma, M)| = O(\varepsilon J_{\sigma,M} M/N) \, .$$

The sampling is successful with probability at least $1 - n^{-3}$ if $p = O(\varepsilon^{-1} \log n)$.

The last assertion in the claim follows from Chernoff bounds. Note that our bound (7) on p is satisfactory, in virtue of the claim.

Summing up the errors $|\hat{A}(\sigma, M) - A(\sigma, M)|$, $|\hat{B}(\sigma, M) - B(\sigma, M)|$ over all M gives us the following assertion: With probability at least $1 - n^{-2}$, uniformly for all σ,

$$|f(\sigma) - \mathrm{reg}_\pi(\sigma)| \le \varepsilon N^{-1} d_{\mathrm{SF}}(\pi, \sigma) \le 2\varepsilon \mathrm{dist}(\pi, \sigma) \, ,$$

where the last inequality is by Diaconis and Graham (1977). This concludes the proof. ■

Algorithm 2 summarizes our specific ε-SRRA construction for LRPP. Note that by the choice of the sample size p, the number of preference queries needed for computing f is $O(\varepsilon^{-3} n \log^4 n)$. Observe that given a pivot π, our LRPP-SRRA construction is simple to implement, and is (obviously) computationally efficient.

Algorithm 2 SRRA for LRPP

Input: V, $\mathcal{C}$, a pivot $\pi \in \mathcal{C}$, estimation parameter $\epsilon \in (0, 1/5)$

```
p ← O(ε^{-3} log^3 n)
for u ∈ V do
    for i = 0, 1, ..., ⌈log n⌉ do
        I_{u,i} ← {v : (2^i − 1)p < |π(u) − π(v)| < 2^{i+1} p}
        for t = 1, ..., p do
            v_{u,i,t} ← a uniformly and independently sampled alternative from I_{u,i}
        end for
    end for
end for
return f : C ⟶ ℝ, defined by
```

$$f(\sigma) = \sum_{u \in V} \sum_{i=0}^{\lceil \log n \rceil} \frac{|I_{u,i}|}{p} \sum_{t=1}^{p} \left(\mathrm{cost}_{u,v_{u,i,t}}(\sigma) - \mathrm{cost}_{u,v_{u,i,t}}(\pi) \right) ,$$

We can now combine these LRPP ε-SRRA constructions in Algorithm 1, the SRRA's method meta-algorithm defined in Corollary 4. This provides the following bound on the number of preference queries.

Corollary 13 *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 $O\left(\varepsilon^{-3} n \log^5 n\right)$. The algorithm succeeds with probability at least $1 - n^{-2}$.*

Corollary 13 tells us that the method of SRRA provides 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, we saw in Section 4.2 that any known active learning results that used bounded disagreement coefficient and VC dimension arguments only guaranteed a query complexity of $O(n\nu^{-1})$, 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 $|\pi^*(u) - \pi^*(v)|^{-\rho}$. Namely, closer pairs of item in the ground truth permutation are more prone to confuse a human labeler. The resulting noise is $\nu = n^{-\rho}$ for some $\rho > 0$. (Note, however, our work does not assume Bayesian noise, and we present this scenario for illustration purposes only.)

In terms of query complexity it turns out that our bound provides only a slight improvement over the divide-and-conquer active learning algorithm for LRPP of Ailon (2012).

Specifically, we improve the dependency on ε from ε^{-6} to ε^{-3}. Although our method provides only a minor improvement it still defines the current state-of-the-art for query efficient LRPP. However, more importantly it defines the first query-efficient LRPP algorithm that is applicable over arbitrary set of permutations, $\mathcal{C} \subseteq V!$. We take advantage of this fact in the Section 6.1, where we instantiate ε-SRRA's for the set of permutations induced by half-spaces in $\mathbb{R}^d$.

5. Application #2: Clustering with Side Information

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 using 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; Balcan et al., 2009; 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; Cesa-Bianchi et al., 2012). 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) studies active learning for clustering using the geometric input structure. Unlike our setting, they assume either no noise or Bayesian noise.

5.1 Problem Definition

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), 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.[12]

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. (For example, 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 a disjoint cover $\{V_1, \ldots, V_k\}$ of V. We take $\mathcal{D}$ 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 $\mathrm{dist}(h,h') = N^{-1} \sum_{(u,v)\in\mathcal{X}} \mathbf{1}_{h(u,v)\neq h'(u,v)}$.

5.2 The Ineffectiveness of Using Disagreement Coefficient Arguments

We demonstrate once again the weakness of a disagreement coefficient approach. It is easy to verify 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 $\mathrm{dist}(h,h') = O(1/\max\{|V_i|\})$ which for a fixed $k = o(n)$ matches $O(1/n)$. 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 would need $\Omega(n^2)$ preference labels to obtain an (ε,μ)-SRRA for any meaningful pair (ε,μ). This is uninformative as the cardinality of $\mathcal{X}$ is $O(n^2)$. As in the problem discussed in Section 4, this can be improved using Remark 4 to $\Omega(n\nu^{-1})$, which tends to quadratic in n as ν becomes smaller.

We note that also here, uniform sampling is doomed to query the entire pool; this can be easily shown using similar arguments to the ones appear in Ailon (2012, Section 2.4). We next show how to construct more useful SRRA's for the problem, for arbitrarily small ν.

12. 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.

5.3 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 with respect to their sizes 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

$$q = c_2 \max\left\{\varepsilon^{-2}k^2, \varepsilon^{-3}k\right\} \log n \tag{12}$$

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 $h' \in \mathcal{C}$,

$$f(h') = \sum_{i=1}^{k} \frac{|V_i|}{q} \sum_{u \in V_i} \sum_{v \in S_{ui}} f_{u,v}(h') + 2 \sum_{i=1}^{k} \sum_{u \in V_i} \sum_{j=i+1}^{k} \frac{|V_j|}{q} \sum_{v \in S_{uj}} f_{u,v}(h') \;, \tag{13}$$

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

Theorem 14 *With probability at least $1 - n^{-3}$ the function f is an ε-SRRA with respect to h.*

Consider another k-clustering $h' \in \mathcal{C}$, with corresponding partitioning $\{V'_1, \ldots, V'_k\}$ of V. We can write $\mathrm{dist}(h, h')$ as

$$\mathrm{dist}(h, h') = \sum_{(u,v) \in \mathcal{X}} \mathrm{dist}_{u,v}(h, h') \,,$$

where $\mathrm{dist}_{u,v}(h, h') = N^{-1}(\mathbf{1}_{h'(u,v)=1}\mathbf{1}_{h(u,v)=0} + \mathbf{1}_{h(u,v)=1}\mathbf{1}_{h'(u,v)=0})$.

Let n_i denote $|V_i|$, and recall that $n_1 \geq n_2 \geq \cdots \geq n_k$. In what follows, we remove the subscript in reg_h and rename it reg (h is held fixed). The function $\mathrm{reg}(h')$ will now be written as:

$$\mathrm{reg}(h') = \sum_{i=1}^{k} \sum_{u \in V_i} \left(\sum_{v \in V_i \setminus \{u\}} \mathrm{reg}_{u,v}(h') + 2 \sum_{j=i+1}^{k} \sum_{v \in V_j} \mathrm{reg}_{u,v}(h') \right) \,,$$

where

$$\mathrm{reg}_{u,v}(h') = \mathrm{cost}_{u,v}(h') - \mathrm{cost}_{u,v}(h) \;.$$

Clearly for each h' it holds that $f(h')$ from (13) is an unbiased estimator of $\mathrm{reg}(h')$. We now analyze its error. For each $i, j \in [k]$ let V_{ij} denote $V_i \cap V'_j$. This captures exactly the set of elements in the i'th cluster in h and the j'th cluster in h'. The distance $\mathrm{dist}(h, h')$ can be written as follows:

$$\mathrm{dist}(h, h') = N^{-1} \left(\sum_{i=1}^{k} \sum_{j=1}^{k} |V_{ij} \times (V_i \setminus V_{ij})| + 2 \sum_{j=1}^{k} \sum_{1 \leq i_1 < i_2 \leq k} |V_{i_1 j} \times V_{i_2 j}| \right) \;. \tag{14}$$

We call each Cartesian set product in (14) a *distance contributing rectangle*. Note that unless a pair (u,v) appears in one of the distance contributing rectangles, we have $\mathrm{reg}_{u,v}(h') = f_{u,v}(h') = 0$. Hence we can decompose $\mathrm{reg}(h')$ and $f(h')$ in correspondence with the distance contributing rectangles, as follows:

$$\mathrm{reg}(h') = \sum_{i=1}^{k}\sum_{j=1}^{k} G_{i,j}(h') + 2\sum_{j=1}^{k}\sum_{1\le i_1<i_2\le k} G_{i_1,i_2,j}(h')\,, \tag{15}$$

$$f(h') = \sum_{i=1}^{k}\sum_{j=1}^{k} F_{i,j}(h') + 2\sum_{j=1}^{k}\sum_{1\le i_1<i_2\le k} F_{i_1,i_2,j}(h')\,, \tag{16}$$

where

$$\begin{aligned} G_{i,j}(h') &= \sum_{u\in V_{ij}}\sum_{v\in V_i\setminus V_{ij}} \mathrm{reg}_{u,v}(h')\,,\\ F_{i,j}(h') &= \frac{|V_i|}{q}\sum_{u\in V_{ij}}\sum_{v\in (V_i\setminus V_{ij})\cap S_{ui}} f_{u,v}(h')\,, \qquad (17)\\ G_{i_1,i_2,j}(h') &= \sum_{u\in V_{i_1 j}}\sum_{v\in V_{i_2 j}} f_{u,v}(h')\,,\\ F_{i_1,i_2,j}(h') &= \frac{|V_{i_2}|}{q}\sum_{u\in V_{i_1 j}}\sum_{v\in V_{i_2 j}\cap S_{ui_2}} f_{u,v}(h')\,. \qquad (18) \end{aligned}$$

(Note that the S_{ui}'s are multisets, and the inner sums in (17) and (18) may count elements multiple times.)

Lemma 15 *With probability at least $1-n^{-3}$, the following holds simultaneously for all $h'\in\mathcal{C}$ and all $i,j\in[k]$:*

$$|G_{i,j}(h') - F_{i,j}(h')| \le \varepsilon N^{-1}\cdot|V_{ij}\times(V_i\setminus V_{ij})|\,. \tag{19}$$

Proof The predicate (19) (for a given i,j) depends only on the set $V_{ij}=V_i\cap V'_j$. Given a subset $B\subseteq V_i$, we say that h' (i,j)-realizes B if $V_{ij}=B$.

Now fix i,j and $B\subseteq V_i$. Assume h' (i,j)-realizes B. Let $\beta=|B|$ and $\gamma=|V_i|$. Consider the random variable $F_{i,j}(h')$. Think of the sample S_{ui} as a sequence $S_{ui}(1),\ldots,S_{ui}(q)$, where each $S_{ui}(s)$ is chosen uniformly at random from V_i for $s=1,\ldots,q$. We can now rewrite $F_{i,j}(h')$ as follows:

$$F_{i,j}(h') = \frac{\gamma}{q}\sum_{u\in B}\sum_{s=1}^{q} Z\big(S_{ui}(s)\big)\,,$$

where

$$Z(v) = \begin{cases} f_{u,v}(h') & v\in V_i\setminus V_{ij}\\ 0 & \text{otherwise}\end{cases}\,.$$

For all $s = 1, \dots q$ the random variable $Z\big(S_{ui}(s)\big)$ is bounded by $2N^{-1}$ almost surely, and its moments satisfy:

$$E\left[Z\big(S_{ui}(s)\big)\right] = \frac{1}{\gamma} \sum_{v \in (V_i \setminus V_{ij})} f_{u,v}(h')\,,$$
$$E\left[Z\big(S_{ui}(s)\big)^2\right] \leq \frac{4N^{-2}(\gamma-\beta)}{\gamma}\ .$$

From this we conclude using Bernstein inequality that for any $t \leq 6\,N^{-1}\beta(\gamma-\beta)$,

$$\Pr\left[\left|F_{i,j}(h') - G_{i,j}(h')\right| \geq t\right] \leq \exp\left\{-\frac{qt^2}{16\gamma\beta(\gamma-\beta)N^{-2}}\right\}\,.$$

Plugging in $t = \varepsilon N^{-1}\beta(\gamma-\beta)$, we conclude

$$\Pr\left[\left|F_{i,j}(h') - G_{i,j}(h')\right| \geq \varepsilon N^{-1}\beta(\gamma-\beta)\right] \leq \exp\left\{-\frac{q\varepsilon^2\beta(\gamma-\beta)}{16\gamma}\right\}\,.$$

Now note that the number of possible sets $B \subseteq V_i$ of size β is at most $n^{\min\{\beta,\gamma-\beta\}}$. Using union bound and recalling our choice of q, the lemma follows. ■

Proving the following is more involved.

Lemma 16 *With probability at least $1 - n^{-3}$, the following holds uniformly for all $h' \in \mathcal{C}$ and for all $i_1, i_2, j \in [k]$ with $i_1 < i_2$:*

$$|F_{i_1,i_2,j}(h') - G_{i_1,i_2,j}(h')| \leq \varepsilon N^{-1} \max\left\{|V_{i_1 j} \times V_{i_2 j}|, \frac{|V_{i_1 j} \times (V_{i_1} \setminus V_{i_1 j})|}{k}, \frac{|V_{i_2 j} \times (V_{i_2} \setminus V_{i_2 j})|}{k}\right\}\,. \tag{20}$$

Proof The predicate (20) (for a given i_1, i_2, j) depends only on the sets $V_{i_1 j} = V_{i_1} \cap V'_j$ and $V_{i_2 j} = V_{i_2} \cap V'_j$. Given subsets $B_1 \subseteq V_{i_1}$ and $B_2 \subseteq V_{i_2}$, we say that h' (i_1, i_2, j)-realizes (B_1, B_2) if $V_{i_1 j} = B_1$ and $V_{i_2 j} = B_2$.

We now fix $i_1 < i_2, j$ and $B_1 \subseteq V_{i_1}$, $B_2 \subseteq V_{i_2}$. Assume h' (i_1, i_2, j)-realizes (B_1, B_2). For brevity, denote $\beta_\iota = |B_\iota|$ and $\gamma_\iota = |V_{i_\iota}|$ for $\iota = 1, 2$. Using Bernstein inequality as in Lemma 15, we conclude the following two inequalities:

$$\Pr\left[\left|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')\right| > t\right] \leq \exp\left\{-\frac{c_3 t^2 q}{\beta_1\beta_2\gamma_2 N^{-2}}\right\} \tag{21}$$

for any t in the range $\left[0, N^{-1}\beta_1\beta_2\right]$, and some global $c_3 > 0$. For t in the range $(N^{-1}\beta_1\beta_2, \infty)$ and some global c_4 we have

$$\Pr\left[\left|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')\right| > t\right] \leq \exp\left\{-\frac{c_4 t q}{\gamma_2 N^{-1}}\right\}\ . \tag{22}$$

Consider the following three cases.

1. $\beta_1\beta_2 \geq \max\{\beta_1(\gamma_1-\beta_1)/k, \beta_2(\gamma_2-\beta_2)/k\}$. Hence, $\beta_1 \geq (\gamma_2-\beta_2)/k, \beta_2 \geq (\gamma_1-\beta_1)/k$. In this case, we can plug $t = \varepsilon N^{-1}\beta_1\beta_2$ in (21) to get

$$\Pr\left[\left|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')\right| > \varepsilon N^{-1}\beta_1\beta_2\right] \leq \exp\left\{-\frac{c_3\varepsilon^2\beta_1\beta_2 q}{\gamma_2}\right\} . \tag{23}$$

Consider two subcases: (i) If $\beta_2 \geq \gamma_2/2$ then the RHS of (23) is at most $\exp\left\{-\frac{c_3\varepsilon^2\beta_1 q}{2}\right\}$. The number of possible subsets B_1, B_2 of sizes β_1, β_2 respectively is clearly at most $n^{\beta_1+(\gamma_2-\beta_2)} \leq n^{\beta_1+k\beta_1}$. Therefore, as long as $q = O(\varepsilon^{-2}k\log n)$ then with probability at least $1-n^{-6}$ this case is taken care of in the following sense: Simultaneously for all $j, i_1 < i_2$, all possible $\beta_1 \leq \gamma_1 = |V_{i_1}|$, $\beta_2 \leq \gamma_2 = |V_{i_2}|$ satisfying the assumptions and for all $B_1 \subseteq V_{i_1 j}, B_2 \subseteq V_{i_2 j}$ of sizes β_1, β_2 respectively and for all h' (i_1, i_2, j)-realizing (B_1, B_2) we have that $|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')| \leq \varepsilon\beta_1\beta_2$.

(ii) If $\beta_2 < \gamma_2/2$ then by our assumption, $\beta_1 \geq \gamma_2/2k$. Hence the RHS of (23) is at most $\exp\left\{-\frac{c_3\varepsilon^2\beta_2 q}{2k}\right\}$. The number of sets B_1, B_2 of sizes β_1, β_2 respectively is clearly at most $n^{(\gamma_1-\beta_1)+\beta_2} \leq n^{\beta_2(1+k)}$. Therefore, as long as $q = O(\varepsilon^{-2}k^2\log n)$ then with probability at least $1-n^{-6}$ this case is taken care of in the following sense: Simultaneously for all $j, i_1 < i_2$, all possible $\beta_1 < \gamma_1 = |V_{i_1}|$, $\beta_2 < \gamma_2 = |V_{i_2}|$ satisfying the assumptions and for all $B_1 \subseteq V_{i_1 j}, B_2 \subseteq V_{i_2 j}$ of sizes β_1, β_2 respectively and for all h' (i_1, i_2, j)-realizing (B_1, B_2) we have that $|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')| \leq \varepsilon\beta_1\beta_2$.

The requirement $q = O(\varepsilon^{-2}k\log n)$ is satisfied by our choice, Equation 12.

2. $\beta_2(\gamma_2-\beta_2)/k \geq \max\{\beta_1\beta_2, \beta_1(\gamma_1-\beta_1)/k\}$. We consider two subcases.

   (a) $\varepsilon\beta_2(\gamma_2-\beta_2)/k \leq \beta_1\beta_2$. Using (21), we get

$$\Pr[|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')| > \varepsilon N^{-1}\beta_2(\gamma_2-\beta_2)/k] \leq \exp\left\{-\frac{c_3\varepsilon^2\beta_2(\gamma_2-\beta_2)^2 q}{k^2\beta_1\gamma_2}\right\} . \tag{24}$$

   Again consider two subcases. (i) $\beta_2 \leq \gamma_2/2$. In this case we conclude from Equation 24

$$\Pr[|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')| > \varepsilon N^{-1}\beta_2(\gamma_2-\beta_2)/k] \leq \exp\left\{-\frac{c_3\varepsilon^2\beta_2\gamma_2 q}{4k^2\beta_1}\right\} . \tag{25}$$

   Now note that by our assumption

$$\beta_1 \leq (\gamma_2-\beta_2)/k \leq \gamma_2/k \leq \gamma_1/k , \tag{26}$$

   the last inequality is in virtue of our assumption $\gamma_1 \geq \gamma_2$. Also by assumption,

$$\beta_1 \leq \beta_2(\gamma_2-\beta_2)/(\gamma_1-\beta_1) \leq \beta_2\gamma_2/(\gamma_1-\beta_1) . \tag{27}$$

   Plugging (26) in the RHS of (27), we conclude that

$$\beta_1 \leq \beta_2\gamma_2/(\gamma_1(1-1/k)) \leq 2\beta_2\gamma_2/\gamma_1 \leq 2\beta_2 .$$

   From here we conclude that the RHS of (25) is at most $\exp\left\{-\frac{c_3\varepsilon^2\gamma_2 q}{8k^2}\right\}$.

The number of sets B_1, B_2 of sizes β_1, β_2 respectively is clearly at most $n^{\beta_1+\beta_2} \leq n^{2\beta_2+\beta_2} \leq n^{3\gamma_2}$. Hence, as long as $q = O(\varepsilon^{-2}k^2 \log n)$ (satisfied by our assumption), with probability at least $1 - n^{-6}$ simultaneously for all $j, i_1 < i_2$, all possible $\beta_1 \leq \gamma_1 = |V_{i_1}|$, $\beta_2 \leq \gamma_2 = |V_{i_2}|$ satisfying the assumptions and for all $B_1 \subseteq V_{i_1 j}, B_2 \subseteq V_{i_2 j}$ of sizes β_1, β_2 respectively and for all h' (i_1, i_2, j)-realizing (B_1, B_2) we have that $|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')| \leq \varepsilon\beta_2(\gamma_2 - \beta_2)/k$. In the second subcase (ii) $\beta_2 > \gamma_2/2$. The RHS of (24) is at most $\exp\left\{-\frac{2c_3\varepsilon^2(\gamma_2-\beta_2)^2 q}{k^2\beta_1}\right\}$. By our assumption, $(\gamma_2-\beta_2)/(k\beta_1) \geq 1$, hence this is at most $\exp\left\{-\frac{2c_3\varepsilon^2(\gamma_2-\beta_2)q}{k}\right\}$. The number of sets B_1, B_2 of sizes β_1, β_2 respectively is clearly at most $n^{\beta_1+(\gamma_2-\beta_2)} \leq n^{(\gamma_2-\beta_2)/k+(\gamma_2-\beta_2)} \leq n^{2(\gamma_2-\beta_2)}$. Therefore, as long as $q = O(\varepsilon^{-2}k \log n)$ (satisfied by our assumption), then with probability at least $1 - n^{-6}$, using a similar counting and union bound argument as above, this case is taken care of in the sense that: $|G_{i_1,i_2,j}(h') - G_{i_1,i_2,j}(h')| \leq \varepsilon\beta_2(\gamma_2 - \beta_2)/k$.

(b) $\varepsilon\beta_2(\gamma_2 - \beta_2)/k > \beta_1\beta_2$. We now use (22) to conclude

$$\Pr[|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')| > \varepsilon N^{-1}\beta_2(\gamma_2 - \beta_2)/k] \leq \exp\left\{-\frac{c_4\varepsilon\beta_2(\gamma_2 - \beta_2)q}{k\gamma_2}\right\} . \tag{28}$$

We again consider the cases (i) $\beta_2 \leq \gamma_2/2$ and (ii) $\beta_2 > \gamma_2/2$ as above. In (i), we get that the RHS of (28) is at most $\exp\left\{-\frac{c_4\varepsilon\beta_2 q}{2k}\right\}$. Now notice that by our assumptions,

$$\beta_1 < \varepsilon(\gamma_2 - \beta_2)/k \leq \gamma_2/2 \leq \gamma_1/2 . \tag{29}$$

Also by our assumptions, $\beta_1 < \beta_2(\gamma_2 - \beta_2)/(\gamma_1 - \beta_1)$, which by (29) is at most $2\beta_2\gamma_2/\gamma_1 \leq 2\beta_2$. Hence the number of possibilities for B_1, B_2 is at most $n^{\beta_1+\beta_2} \leq n^{3\beta_2}$. In (ii), we get that the RHS of (28) is at most $\exp\left\{-\frac{c_4\varepsilon(\gamma_2-\beta_2)q}{2k}\right\}$, and the number of possibilities for B_1, B_2 is at most $n^{\beta_1+(\gamma_2-\beta_2)}$ which is bounded by $n^{2(\gamma_2-\beta_2)}$ by our assumptions. For both (i) and (ii) taking $q = O(\varepsilon^{-1}k \log n)$ ensures with probability at least $1-n^{-6}$, using a similar counting and union bounding argument as above, case (b) is taken care of in the sense that: $|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')| \leq \varepsilon N^{-1}\beta_2(\gamma_2 - \beta_2)/k$.

3. $\beta_1(\gamma_1 - \beta_1)/k \geq \max\{\beta_1\beta_2, \beta_2(\gamma_2 - \beta_2)/k\}$. We consider two subcases.

(a) $\varepsilon\beta_1(\gamma_1 - \beta_1)/k \leq \beta_1\beta_2$. Using (21), we get

$$\Pr\left[|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')| > \varepsilon N^{-1}\beta_1(\gamma_1 - \beta_1)/k\right] \leq \exp\left\{-\frac{c_3\varepsilon^2\beta_1(\gamma_1 - \beta_1)^2 q}{k^2\beta_2\gamma_2}\right\} . \tag{30}$$

As before, consider case (i) in which $\beta_2 \leq \gamma_2/2$ and (ii) in which $\beta_2 > \gamma_2/2$. For case (i), we use the fact that $\beta_1(\gamma_1 - \beta_1) \geq \beta_2(\gamma_2 - \beta_2)$ by assumption and notice that the RHS of (30) is at most $\exp\left\{-\frac{c_3\varepsilon^2\beta_2(\gamma_2-\beta_2)(\gamma_1-\beta_1)q}{k^2\beta_2\gamma_2}\right\}$. This is hence at most $\exp\left\{-\frac{c_3\varepsilon^2(\gamma_1-\beta_1)q}{2k^2}\right\}$. The number of possibilities of B_1, B_2 of sizes β_1, β_2 is clearly at most

$n^{(\gamma_1-\beta_1)+\beta_2} \leq n^{(\gamma_1-\beta_1)+(\gamma_1-\beta_1)/k} \leq n^{2(\gamma_1-\beta_1)}$. From this we conclude that $q = O(\varepsilon^{-2}k^2 \log n)$ suffices for this case. For case (ii), we bound the RHS of (30) by $\exp\left\{-\frac{c_3\varepsilon^2\beta_1(\gamma_1-\beta_1)^2 q}{2k^2\beta_2^2}\right\}$. Using the assumption that $(\gamma_1-\beta_1)/\beta_2 \geq k$, the latter expression is upper bounded by $\exp\left\{-\frac{c_3\varepsilon^2\beta_1 q}{2}\right\}$. Again by our assumptions,

$$\beta_1 \geq \beta_2(\gamma_2-\beta_2)/(\gamma_1-\beta_1) \geq (\varepsilon(\gamma_1-\beta_1)/k)(\gamma_2-\beta_2)/(\gamma_1-\beta_1) = \varepsilon(\gamma_2-\beta_2)/k \ . \tag{31}$$

The number of possibilities of B_1, B_2 of sizes β_1, β_2 is clearly at most $n^{\beta_1+(\gamma_2-\beta_2)}$ which by (31) is bounded by $n^{\beta_1+k\beta_1/\varepsilon} \leq n^{2k\beta_1/\varepsilon}$. From this we conclude that as long as $q = O(\varepsilon^{-3}k\log n)$ (satisfied by our choice), this case is taken care of in sense repeatedly explained above.

(b) $\varepsilon\beta_1(\gamma_1-\beta_1)/k > \beta_1\beta_2$. Using (22), we get

$$\Pr\left[\left|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')\right| > \varepsilon N^{-1}\beta_1(\gamma_1-\beta_1)/k\right] \leq \exp\left\{-\frac{c_4\varepsilon\beta_1(\gamma_1-\beta_1)q}{k\gamma_2}\right\} \ . \tag{32}$$

We consider two sub-cases, (i) $\beta_1 \leq \gamma_1/2$ and (ii) $\beta_1 > \gamma_1/2$. In case (i), we have that

$$\begin{aligned}\frac{\beta_1(\gamma_1-\beta_1)}{\gamma_2} &= \frac{1}{2}\frac{\beta_1(\gamma_1-\beta_1)}{\gamma_2} + \frac{1}{2}\frac{\beta_1(\gamma_1-\beta_1)}{\gamma_2} \\ &\geq \frac{1}{2}\frac{\beta_1\gamma_1}{2\gamma_2} + \frac{1}{2}\frac{\beta_2(\gamma_2-\beta_2)}{\gamma_2} \\ &\geq \beta_1/4 + \min\{\beta_2, \gamma_2-\beta_2\}/2 \ .\end{aligned}$$

(The last step used $\gamma_1 \geq \gamma_2$.) Hence, the RHS of (32) is bounded above by

$$\exp\left\{-\frac{c_4\varepsilon q(\beta_1/4 + \min\{\beta_2, \gamma_2-\beta_2\}/2)}{k}\right\} \ .$$

The number of possibilities of B_1, B_2 of sizes β_1, β_2 is clearly at most $n^{\beta_1+\min\{\beta_2,\gamma_2-\beta_2\}}$, hence as long as $q = O(\varepsilon^{-1}k\log n)$ (satisfied by our choice), this case is taken care of in the sense repeatedly explained above. In case (ii), we can upper bound the RHS of (32) by $\exp\left\{-\frac{c_4\varepsilon\gamma_1(\gamma_1-\beta_1)q}{2k\gamma_2}\right\} \geq \exp\left\{-\frac{c_4\varepsilon(\gamma_1-\beta_1)q}{2k}\right\}$. The number of possibilities of B_1, B_2 of sizes β_1, β_2 is clearly at most $n^{(\gamma_1-\beta_1)+\beta_2}$ which, using our assumptions, is bounded above by $n^{(\gamma_1-\beta_1)+(\gamma_1-\beta_1)/k} \leq n^{2(\gamma_1-\beta_1)}$. Hence, as long as $q = O(\varepsilon^{-1}k\log n)$, this case is taken care of in the sense repeatedly explained above.

This concludes the proof of the lemma. ■

As a consequence, we get the following:

Lemma 17 *with probability at least* $1-n^{-3}$*, the following holds simultaneously for all* k*-clusterings* $\mathcal{C}'$*:* $|\mathrm{reg}(h') - f(h')| \leq 5\varepsilon\mathrm{dist}(h', h)$.

Proof By the triangle inequality,

$$|\mathrm{reg}(h') - f(h')| \leq \sum_{i=1}^{k}\sum_{j=1}^{k}|G_{i,j}(h') - F_{i,j}(h')| + 2\sum_{j=1}^{k}\sum_{1\leq i_1<i_2\leq k}|G_{i_1,i_2,j}(h') - F_{i_1,i_2,j}(h')|\,.$$

Using (15)-(16), then Lemmas 15 and 16 (assuming success of a high probability event), and rearranging the sum and finally using (14), we get:

$$\begin{aligned}
|\mathrm{reg}(h') - f(h')| \leq & \sum_{i=1}^{k}\sum_{j=1}^{k}\varepsilon N^{-1}|V_{ij}\times(V_i\setminus V_{ij})| \\
& + 2\varepsilon N^{-1}\sum_{j=1}^{k}\sum_{i_1<i_2}\left(|V_{i_1j}\times V_{i_2j}| + k^{-1}|V_{i_1j}\times(V_{i_1}\setminus V_{i_1j})| + k^{-1}|V_{i_2j}\times(V_{i_2}\setminus V_{i_2j})|\right) \\
\leq & \sum_{i=1}^{k}\sum_{j=1}^{k}\varepsilon N^{-1}|V_{ij}\times(V_i\setminus V_{ij})| + 2\varepsilon N^{-1}\sum_{j=1}^{k}\sum_{i_1<i_2}|V_{i_1j}\times V_{i_2j}| \\
& + 2\varepsilon N^{-1}\sum_{j=1}^{k}\sum_{i_1<i_2}^{k}k^{-1}|V_{i_1j}\times(V_{i_1}\setminus V_{i_1j})| + 2\varepsilon N^{-1}\sum_{j=1}^{k}\sum_{i_1<i_2}^{k}k^{-1}|V_{i_2j}\times(V_{i_2}\setminus V_{i_2j})| \\
\leq & \sum_{i=1}^{k}\sum_{j=1}^{k}\varepsilon N^{-1}|V_{ij}\times(V_i\setminus V_{ij})| + 2\varepsilon N^{-1}\sum_{j=1}^{k}\sum_{i_1<i_2}|V_{i_1j}\times V_{i_2j}| \\
& + 2\varepsilon N^{-1}\sum_{j=1}^{k}\sum_{i_1=1}^{k}kk^{-1}|V_{i_1j}\times(V_{i_1}\setminus V_{i_1j})| + 2\varepsilon N^{-1}\sum_{j=1}^{k}\sum_{i_2=1}^{k}kk^{-1}|V_{i_2j}\times(V_{i_2}\setminus V_{i_2j})| \\
& \leq 5\varepsilon N^{-1}\sum_{i=1}^{k}\sum_{j=1}^{k}|V_{ij}\times(V_i\setminus V_{ij})| + \varepsilon N^{-1}\sum_{j=1}^{k}\sum_{i_1<i_2}|V_{i_1j}\times V_{i_2j}| \leq 5\varepsilon\mathrm{dist}(h,h')\,,
\end{aligned}$$

as required. ■

We conclude that f is an ε-SRRA estimator. Its construction pseudo code is presented for convenience in Algorithm 3. Clearly the number of label queries required for obtaining this ε-SRRA estimator is $O(n\max\{\varepsilon^{-2}k^3,\varepsilon^{-3}k^2\}\log n)$. Combining Theorem 14 with this bound and the iterative algorithm described in Corollary 4 (Algorithm 1), we obtain the following:

Corollary 18 *There exists an active learning algorithm for obtaining a solution $h\in\mathcal{C}$ for semi-supervised k-clustering with $\mathrm{er}_{\mathcal{D}}(h)\leq(1+O(\varepsilon))\,\nu$ 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 $1-n^{-2}$.*

We do not believe the ε^{-3} factor in the corollary is tight, and speculate that it should be reduced to ε^{-2} using more advanced measure concentration tools. Note that we assume k is fixed. Indeed, in practice, k is often taken to be constant (or at most $o(n)$). Thus, the sample complexity of our active learning method using these direct SRRA constructions is almost linear in n. As in the case of Corollary 13 and the ensuing discussion around LRPP, the result in Corollary 18 significantly beats known active learning results depending only on disagreement coefficient and VC dimension bounds, for small ν.

Algorithm 3 SRRA for Semi-Supervised k-Clustering

Input: V, k, $\mathcal{C}$, a pivot $h = \{V_i\}_{i=1}^k \in \mathcal{C}$, estimation parameter $\epsilon \in (0, 1/5)$

$q \leftarrow O(\max\left\{\varepsilon^{-2}k^2, \varepsilon^{-3}k\right\}\log n)$
Index the clusters of h such that $|V_1| \geq |V_2| \geq \ldots \geq |V_k|$
for $u \in V_i,\ i = 1, \ldots, k$ **do**
  for $j = i, \ldots, k$ **do**
    $S_{u,j} \leftarrow$ sample q elements from V_j independently and uniformly with repetitions
  end for
end for
return $f : \mathcal{C} \longrightarrow \mathbb{R}$, defined by

$$f(h') = \sum_{i=1}^{k} \frac{|V_i|}{q} \sum_{u \in V_i} \sum_{v \in S_{ui}} \left(\mathrm{cost}_{u,v}(h') - \mathrm{cost}_{u,v}(h)\right) + 2\sum_{i=1}^{k} \sum_{u \in V_i} \sum_{j=i+1}^{k} \frac{|V_j|}{q} \sum_{v \in S_{uj}} \left(\mathrm{cost}_{u,v}(h') - \mathrm{cost}_{u,v}(h)\right)$$

6. Additional Results and Practical Considerations

We discuss two practical extensions of our results.

6.1 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, except that now each point $v \in V$ is associated with a feature vector, which we denote using bold face: $\mathbf{v} \in \mathbb{R}^d$. The concept space $\mathcal{C}$ now consists only of permutations π such that there exists a vector $\mathbf{w}_\pi \in \mathbb{R}^d$ satisfying

$$\pi(u, v) = 1 \iff \langle \mathbf{w}_\pi, \mathbf{u} - \mathbf{v} \rangle > 0 \ . \tag{33}$$

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 $H_{u,v} = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{u} - \mathbf{v} \rangle > 0\}$, whose (closure) supporting hyperplane is $h_{u,v} = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{u} - \mathbf{v} \rangle = 0\}$. Let $\mathcal{H}$ be the collection of these $\binom{n}{2}$ hyperplanes $\{h_{u,v} : (u, v) \in \mathcal{X}\}$.[13] 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 $\mathbb{C}_\pi \subseteq \mathbb{R}^d$ the unique cell corresponding to a permutation $\pi \in \mathcal{C}$.

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

6.1.1 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 $\text{dist}(\sigma, \pi) \leq r$. The corresponding disagreement region $\text{DIS}(\mathcal{B}(\pi, r))$ corresponds to the set of ordered pairs (halfspaces) intersecting $\mathcal{B}(\pi, r)$. In fact, in this case, all these halfspaces have the property that their bounding hyperplane h meets $\mathcal{B}(\pi, r)$. Indeed, if that hyperplane is located outside $\mathcal{B}(\pi, r)$ then there is a clear agreement among all cells (hypeotheses) of $\mathcal{B}(\pi, r)$ with respect to H, as all of them are located at the same side of h. We next show:

Proposition 19 *The measure of* $\text{DIS}\left(\mathcal{B}(\pi, r)\right)$ *in* $\mathcal{D}_{\mathcal{X}}$ *is at most* $8rn$.

Proof By Diaconis and Graham (1977), the Spearman Footrule distance between any two permutations π and σ is at most twice $N\text{dist}(\pi, \sigma)$, where $N = n(n-1)$. Hence, if $\text{dist}(\pi, \sigma)$ is r, then any element u could only swap locations with a set of elements located up to $2rN$ positions away to the right or to the left. This yields a total of $4rN$ 'swap-candidates' for each u. Thus, at most $4rNn$ inversions are possible. Note that each inversion corresponds to a hyperplane (unordered pair) that we cross, and thus the total number of ordered pairs is at most $8rNn$. The probability measure of this set is at most $8rn$, because we assign equal probability of N^{-1} for each possible pair in $\mathcal{X}$. The result follows. ■

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 20 *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 20, and the iterative algorithm described in Corollary 4:

Corollary 21 *There exists an algorithm for obtaining a solution* $\pi \in \mathcal{C}$ *for LRPP in linearly induced permutations in* d *dimensional feature space with* $\text{er}_{\mathcal{D}}(\pi) \leq (1 + O(\varepsilon))\,\nu$ *with total query complexity of*

$$O\Big(\varepsilon^{-2}nd\log^3 n + n\varepsilon^{-2}(\log^2 n)\left(\log(\delta^{-1}\log n)\right)\Big). \tag{34}$$

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

We compare this bound to that of Corollary 13. For the sake of comparison, assume $\delta = n^{-2}$, so that (34) takes the simpler form of $O\big(\varepsilon^{-2} nd \log^3 n / \log(1/\varepsilon)\big)$. This bound is better than that of Corollary 13 as long as the feature space dimension d is $O(\varepsilon^{-2}\log^2 n)$. For larger dimensions, Corollary 13 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 conclude the discussion with a polynomial time algorithm to construct the disagreement region obtained in this setting.

6.1.2 An Algorithm for Constructing the Disagreement Region

We first recall that the disagreement region $\mathrm{DIS}(\mathcal{B}(\pi, r))$ corresponds to the set of bounding hyperplanes h that meet $\mathcal{B}(\pi, r)$. Our goal is thus to compute this set of hyperplanes, their side of space containing π will then correspond to the desired set of halfspaces.

We next introduce, for the sake of analysis, another (absolute) measure related to our previous definitions: We define the *absolute distance* of a hyperplane h from a point p in a cell $\mathbb{C}_\pi$ of $\mathcal{A}(\mathcal{H})$ to be the smallest number of hyperplanes, whose removal makes h visible to p, implying that h appears on the boundary of the cell containing p in the arrangement of this subset of hyperplanes. In what follows, and with a slight abuse of notation, we will sometimes refer to the absolute distance of h from a permutation π (instead of a point p in the cell $\mathbb{C}_\pi$).

By the definition of the measure r it follows that when a hyperplane h intersects $\mathcal{B}(\pi, r)$ this implies that the absolute distance between h and any point in $\mathbb{C}_\pi$ is at most $\bar{r} := Nr$, where $N = n(n-1)$. We also observe that, by definition, at least one cell in $\mathcal{B}(\pi, r)$ must have h on its boundary.

In order to compute all hyperplanes meeting $\mathcal{B}(\pi, r)$ we construct this set iteratively. At the first iteration, we compute all hyperplanes with (absolute) distance 0 from π—these hyperplanes are precisely those that define the boundary of $\mathbb{C}_\pi$. Then we remove them from further consideration, and recursively compute the next set of hyperplanes at (absolute) distance 0 from π in the arrangement of the residual hyperplanes. We stop after $\bar{r}$ iterations. By definition, at the ith step, we compute the set of hyperplanes at absolute distance i from π. Clearly, the number of iterations is at most $|\mathcal{H}| = \binom{n}{2}$, so it is sufficient to show that each iteration can be performed in polynomial time.

We thus face the following problem. Given a set of hyperplanes $\mathcal{H}$ and a point p, determine in polynomial time the subset of hyperplanes in $\mathcal{H}$ that define the cell $\mathbb{C}_p$ of p in the arrangement $\mathcal{A}(\mathcal{H})$. This task can easily be done by constructing the entire arrangement $\mathcal{A}(\mathcal{H})$ or even just the cell containing p. Nevertheless, these constructions take time $n^{\Theta(d)}$ (see, e.g., Sharir and Agarwal 1995), and thus may be inefficient for large values of d. We thus present below a simple approach that circumvents the need to construct these structures, our key idea is to use *linear programming solvers*, and is a variant of the analysis in Ezra and Fine (2007).

To this end, fix a hyperplane $h \in \mathcal{H}$. In order to determine whether h appears on the boundary of the cell containing p, we proceed as follows. For each hyperplane $h' \in \mathcal{H} \setminus \{h\}$ we assign the corresponding halfspace H' bounded by h' and containing p, so p lies at the same side (positive or negative) of these hyperplanes (after, e.g., applying a proper

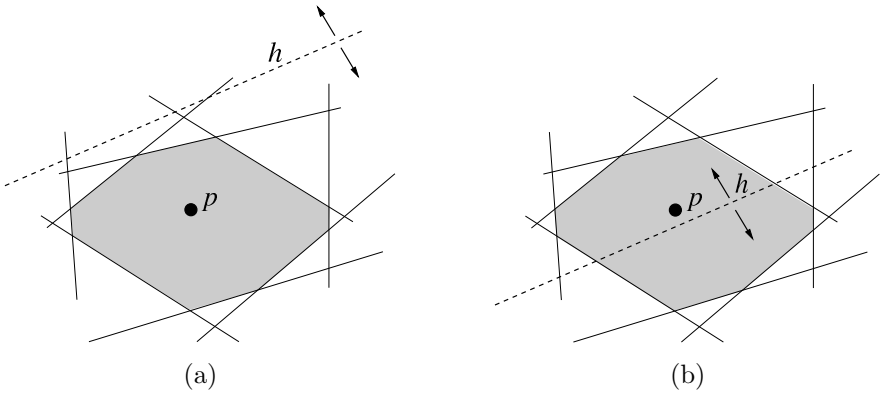

Figure 1: (a) The hyperplane h (depicted by the dashed line in the figure) does not meet the cell containing the point p. (b) The hyperplane h meets that cell. In this case p lies in the portion of that cell located above h.

orientation). Let $\mathcal{H}'$ be the set of these halfspaces. Then we create two additional halfspaces, H^+ and H^- containing the corresponding positive and negative sides of h. We now observe that if h does not appear on the boundary of $\mathbb{C}_p$, then either $\bigcap_{H' \in \mathcal{H}'} H' \bigcap H^+ = \emptyset$ or $\bigcap_{H' \in \mathcal{H}'} H' \bigcap H^- = \emptyset$. Indeed, $\bigcap_{H' \in \mathcal{H}'} H'$ is precisely the cell containing p in $\mathcal{A}(\mathcal{H} \setminus \{h\})$, and if h does not appear on the boundary of $\mathbb{C}_p$ (the cell containing p in $\mathcal{A}(\mathcal{H})$) then we must have that h and $\bigcap_{H' \in \mathcal{H}'} H'$ are disjoint, and thus $\bigcap_{H' \in \mathcal{H}'} H'$ is fully contained in one side of h. If h appears on the boundary of $\mathbb{C}_p$ then both intersections $\bigcap_{H' \in \mathcal{H}'} H' \bigcap H^+$, $\bigcap_{H' \in \mathcal{H}'} H' \bigcap H^-$ must not be empty. In fact, in this case h splits the cell containing p in $\mathcal{A}(\mathcal{H} \setminus \{h\})$ into the two portions $\bigcap_{H' \in \mathcal{H}'} H' \bigcap H^+$, $\bigcap_{H' \in \mathcal{H}'} H' \bigcap H^-$, one of which contains p. See Figure 1 for an illustration. In order to determine if each of these intersections is empty, we construct the corresponding linear programming system, where the constraints are the halfspaces and the objective function is arbitrary, and solve it in polynomial time, using, for example, the ellipsoid method or the interior point method (Grötschel et al., 1988; Karmarkar, 1984; Khačiyan, 1979).

It thus follows that for each hyperplane $h \in H$ we need to solve two linear programming systems in order to determine whether h defines $\mathbb{C}_p$. When we terminate iterating over all $h \in \mathcal{H}$ we have at hand the desired set of hyperplanes of $\mathcal{H}$ that define $\mathbb{C}_p$. Thus the running time of the entire process is polynomial, since at each iteration we spend polynomial time. We have thus obtained:

Theorem 22 *Given a set $\mathcal{H}$ of hyperplanes and a point p in d-space, one can determine in polynomial time the subset of hyperplanes that define the cell in the arrangement $\mathcal{A}(\mathcal{H})$ containing p, where the degree of the polynomial is an absolute constant that does not depend on the dimension d.*

Based on the iterative algorithm presented above, we thus conclude:

Corollary 23 *Given a set $\mathcal{H}$ of hyperplanes in d-space, a permutation π corresponding to a cell $\mathbb{C}_\pi$ in the arrangement $\mathcal{A}(\mathcal{H})$, and a parameter $0 \leq r \leq 1$, one can determine*

in polynomial time the subset of hyperplanes intersecting $\mathcal{B}(\pi, r)$, where the degree of the polynomial is an absolute constant that does not depend on the dimension d.

Remark: We note that since the degree of the polynomial at the running time does not depend on d, we can assume the dimension d is arbitrarily large.

6.2 Convex Relaxations

So far we focused on theoretical ERM aspects only. Doing so, we made no assumptions about the computability of the step $h_i = \operatorname{argmin}_{h' \in \mathcal{C}} f_{h_{i-1}}(h')$ in Corollary 4 (Step 2 in Algorithm 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 $\mathrm{er}_{\mathcal{D}}(\tilde{h})$ over $\tilde{h} \in \widetilde{\mathcal{C}}$, where $\widetilde{\mathcal{C}}$ is a convex set of functions from $\mathcal{X}$ to $\mathbb{R}$. Also assume there is a mapping $\phi : \widetilde{\mathcal{C}} \mapsto \mathcal{C}$ which is used as a "rounding" procedure. For example, in the setting of Section 6.1 the set $\widetilde{\mathcal{C}}$ consists of all vectors $\mathbf{w} \in \mathbb{R}^d$, and the rounding method $\phi : \widetilde{\mathcal{C}} \mapsto \mathcal{C}$ converts $\mathbf{w}$ to a permutation π satisfying (33). When optimizing in $\widetilde{\mathcal{C}}$, one conveniently works with a convex relaxation $\widetilde{\mathrm{er}}_{\mathcal{D}} : \widetilde{\mathcal{C}} \to \mathbb{R}^+$ as surrogate for the discrete loss $\mathrm{er}_{\mathcal{D}}$, defined as follows

$$\widetilde{\mathrm{er}}_{\mathcal{D}}(\tilde{h}) = \mathbf{E}_{(X,Y)\sim\mathcal{D}}\left[\tilde{L}\left(\tilde{h}(X), Y\right)\right] \ . \tag{35}$$

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

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

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 6.1 would be to define for all $\mathbf{w} \in \widetilde{\mathcal{C}}$ 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, optimizing over (35) becomes the famous SVMRank with the hinge loss relaxation (Herbrich et al., 2000; Joachims, 2002):

$$\begin{aligned} \text{Minimize } F(\mathbf{w}, \xi) = \quad & \sum_{u,v} \xi_{u,v} \\ \text{s.t., } \forall u, v, Y(u,v) = 1 : \quad & (\mathbf{u} - \mathbf{v}) \cdot \mathbf{w} \geq 1 - \xi_{u,v} \\ \forall u, v : \quad & \xi_{u,v} \geq 0 \,, \\ & \|\mathbf{w}\| \leq c. \end{aligned}$$

(Note: c is a regularization parameter.)

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, $\widetilde{\mathrm{reg}}_{\tilde{h}} : \widetilde{\mathcal{C}} \mapsto \mathbb{R}^+$ can be efficiently optimized. We now say that $f : \tilde{\mathcal{C}} \mapsto \mathbb{R}^+$ is an (ε, μ)-SRRA with respect to $\tilde{h} \in \widetilde{\mathcal{C}}$ if for all $\tilde{h}' \in \widetilde{\mathcal{C}}$,

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

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

Theorem 24 *Let $\tilde{h}_0, \tilde{h}_1, \tilde{h}_2, \ldots$ be a sequence of hypotheses in $\widetilde{\mathcal{C}}$ such that for all $i \geq 1$, $\tilde{h}_i = \operatorname{argmin}_{\tilde{h}' \in \widetilde{\mathcal{C}}} f_{i-1}(\tilde{h}')$, 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 $\tilde{\nu} = \inf_{\tilde{h} \in \widetilde{C}} \widetilde{\mathrm{er}}_{\mathcal{D}}(\tilde{h})$ 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 turns out that the sampling techniques used for constructing an ε-SRRA from Section 4.3 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. It is easy to confirm that under such bounded-norm setting all arguments of Section 4.3 follows through. The conclusion is that we can solve SVMRank, in polynomial time, to within an error of $(1 + \varepsilon)\tilde{\nu}$ using only $O\big(n \operatorname{poly}(\log n, \varepsilon^{-1})\big)$ preference queries.

6.2.1 Discussion

1. It is worth mentioning the empirically evidence for the success of learning to rank methods that are focused on using relaxed rather than exact constraints on the solution function (see, e.g., Mcfee and Lanckriet, 2010; Long et al., 2010). Another possible empirically viable direction is to make structural assumptions on the preference noise, we refer the reader to the work of Jamieson and Nowak (2011) for a recent result with improved query complexity under certain Bayesian noise assumptions.
2. We note that while disagreement coefficients are brittle, SRRAs are more robust, as demonstrated by the results reported in this paper. Specifically, the application of SRRAs yields efficient solutions to problems on which a disagreement coefficient approach fails, as described in Sections 4.2 and 5.2, and thus we believe SRRAs should be applied more broadly than a stand-alone disagreement coefficient approach.

7. 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 (currently in the margin but gradually moving inside the active learning literature), 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, Nina Balcan, Miroslav Dudik, Ran El-Yaniv, Sariel Har-Peled, John Langford, Rob Schapire, Masashi Sugiyama, and Yair Weiner for helpful discussions. Nir Ailon acknowledges the support of a Marie Curie International Reintegration Grant PIRG07-GA-2010-268403. Esther Ezra acknowledges the support of a National Science Foundation Grant CCF-12-16689.

References

Nir 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.

Nir Ailon, Bernard Chazelle, Seshadhri Comandur, and Ding Liu. Estimating the distance to a monotone function. *Random Struct. Algorithms*, 31(3):371–383, 2007.

Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: Ranking and clustering. *Journal of the ACM*, 55(5):23:1–23:27, October 2008.

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

Francis 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.

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

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

Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman. The true sample complexity of active learning. In *COLT*, pages 45–56, 2008.

Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, pages 1068–1077, 2009.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56:89–113, 2004.

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

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

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 2009.

Alina Beygelzimer, Daniel Hsu, John Langford, and Tong Zhang. Agnostic active learning without constraints. In *NIPS*, 2010.

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

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

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

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

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

Nicolò Cesa-Bianchi, Claudio Gentile, Fabio Vitale, and Giovanni Zappella. Active learning on trees and graphs. In *COLT*, pages 320–332, 2010.

Nicolò Cesa-Bianchi, Claudio Gentile, Fabio Vitale, and Giovanni Zappella. A correlation clustering approach to link classification in signed networks. In *Proceedings of the 25th Annual Conference on Learning Theory. JMLR Workshop and Conference Proceedings*, volume 23 of *JMLR Workshop and Conference Proceedings*, pages 34.1–34.20, 2012.

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

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

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

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

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

Sanjoy Dasgupta, Daniel Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. In *NIPS*, 2007.

Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin, 3rd edition, 2008.

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

Persi Diaconis and R. L. Graham. Spearman's Footrule as a measure of disarray. *Journal of the Royal Statistical Society*, 39(2):262–268, 1977.

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

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

Esther Ezra and Shai Fine. On the cover of convex polyhedra in d-space. Unpublished manuscript, 2007.

Yoav Freund, Sebastian H. Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, September 1997.

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

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2 of *Algorithms and Combinatorics*. Springer-Verlag, 1988.

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

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

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

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

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

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

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin ranking boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, chapter 7, pages 115–132. The MIT Press, 2000.

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

Thorsten Joachims. Optimizing search engines using clickthrough data. In *KDD*, 2002.

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

Narendra Karmarkar. A new polynomial–time algorithm for linear programming. *Combinatorica*, 4:373–395, 1984.

Claire Kenyon-Mathieu and Warren 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.

Leonid Khačiyan. Polynomial algorithm for linear programming. *Soviet Doklady*, 244: 1093–1096, 1979.

Dan Klein, Sepandar D. Kamvar, and Christopher 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.

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

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

Bo Long, Olivier Chapelle, Ya Zhang, Yi Chang, Zhaohui Zheng, and Belle Tseng. Active learning for ranking through expected loss optimization. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–274, 2010.

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

Brian Mcfee and Gert Lanckriet. Metric learning to rank. In *In Proceedings of the 27th Annual International Conference on Machine Learning (ICML)*, 2010.

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

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

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

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

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

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

Micha Sharir and Pankaj K. Agarwal. *Davenport-Schinzel Sequences and Their Geometric Applications*. Cambridge University Press, 1995.

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

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

Konstantin Voevodski, Maria-Florina Balcan, Heiko Röglin, Shang-Hua Teng, and Yu Xia. Active clustering of biological sequences. *Journal of Machine Learning Research*, 13: 203–225, 2012.

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

Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart 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.

Liu Yang, Steve Hanneke, and Jaime G. Carbonell. Bayesian active learning using arbitrary binary valued queries. In *ALT*, pages 50–58, 2010.

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