Journal of Machine Learning Research 13 (2012) 2655-2697 Submitted 12/11; Published 9/12

Selective Sampling and Active Learning from Single and Multiple Teachers

Ofer Dekel OFERD@MICROSOFT.COM
Microsoft Research
One Microsoft Way
Redmond, WA, 98052, USA

Claudio Gentile CLAUDIO.GENTILE@UNINSUBRIA.IT
DiSTA, Università dell'Insubria
via Mazzini 5
21100 Varese, Italy

Karthik Sridharan SKARTHIK@WHARTON.UPENN.EDU
Department of Statistics of the Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA, 19104, USA

Editor: Sanjoy Dasgupta

Abstract

We present a new online learning algorithm in the selective sampling framework, where labels must be actively queried before they are revealed. We prove bounds on the regret of our algorithm and on the number of labels it queries when faced with an adaptive adversarial strategy of generating the instances. Our bounds both generalize and strictly improve over previous bounds in similar settings. Additionally, our selective sampling algorithm can be converted into an efficient statistical active learning algorithm. We extend our algorithm and analysis to the multiple-teacher setting, where the algorithm can choose which subset of teachers to query for each label. Finally, we demonstrate the effectiveness of our techniques on a real-world Internet search problem.

Keywords: online learning, regret, label-efficient, crowdsourcing

1. Introduction

Human-generated labels are expensive. The *active learning* paradigm is built around the idea that we should only acquire labels that actually improve our ability to make accurate predictions. *Online selective sampling* (Cohn et al., 1990; Freund et al., 1997) is an active learning setting that is modeled as a repeated game between a *learner* and an *adversary*. On round t of the game, the adversary presents the learner with an instance $\mathbf{x}_t \in \mathbb{R}^d$ and the learner responds by predicting a binary label $\hat{y}_t \in \{-1,+1\}$. The learner has access to a *teacher*,[1] who knows the correct label for each instance. The learner must now decide whether or not to pay a unit cost and query the teacher for the correct binary label $y_t \in \{-1,+1\}$ from the teacher. If the learner decides to issue a query, he observes the correct label and uses it to improve his future predictions. However, when we analyze the accuracy

1. Most previous publications do not distinguish between the *adversary* and the *teacher*. We make this distinction explicitly and intentionally, in anticipation of the multiple-teacher variant of the problem.

of the learner's predictions, we account for all labels, regardless of whether they were observed by the learner or not. The learner has two conflicting goals: to make accurate predictions and to issue a small number of queries.

To motivate the selective sampling setting, consider an Internet search company that uses online learning techniques to construct a (simplified) search engine. In this case, the instance $\mathbf{x}_t$ represents the pairing of a search-engine query with a candidate web page and the task is to predict whether this pair is a good match or a bad match. Clearly, there is no way to manually label the millions of daily search engine queries along with all of their candidate web pages. Instead, an intelligent mechanism of choosing which instances to label is required. Search engine queries arrive in an online manner and a search engine uses its index of the web to match each query with potential candidate URLs, making this problem well suited for the selective sampling problem setting.

The first part of this paper is devoted to the selective sampling framework described above. In Section 2 we present a selective sampling learning algorithm inspired by known ridge regression algorithms (Hoerl and Kennard, 1970; Lai and Wei, 1982; Vovk, 2001; Azoury and Warmuth, 2001; Cesa-Bianchi et al., 2003, 2005a; Li et al., 2008; Strehl and Littman, 2008; Cavallanti et al., 2009; Cesa-Bianchi et al., 2009). To analyze this algorithm, we adopt the model introduced in Cavallanti et al. (2009), Cesa-Bianchi et al. (2009) and Strehl and Littman (2008), where the adversary may choose arbitrary instances, but the teacher is stochastic and samples each label from an instance-dependent distribution. We evaluate the accuracy of the learner using the game-theoretic notion of *regret*, which measures the extent to which the learner's predictions disagree with the teacher's labels. We prove both an upper bound on the regret and an upper bound on the number of queries issued by the learner.

Our algorithm is an online learning algorithm, designed to incrementally make binary predictions on a sequence of adversarially-generated instances. However, we can also convert our algorithm into an efficient statistical active learning algorithm, which receives a sample of instances from some unknown distribution, queries the teacher for a subset of the labels, and outputs a hypothesis with a small risk. The risk of a hypothesis is its error rate on new instances sampled form the same underlying distribution. We present the details of this conversion in Section 2.5.

In the setting described above, we assumed the learner has access to a single all-knowing teacher. To make things more interesting, we introduce multiple teachers, each with a different area of expertise and a different level of overall competence. On round t, some of the teachers may be experts on $\mathbf{x}_t$ while others may not be. A teacher who is an expert on $\mathbf{x}_t$ is likely to provide the correct label, while a teacher who isn't may give the wrong label. To make this setting as realistic as possible, we assume that the areas of expertise and the overall competence levels of the different teachers are unknown to the learner, and any characterization of a teacher must be inferred from the observed labels.

On round t, the learner receives the instance $\mathbf{x}_t$ from the adversary and makes the binary prediction $\hat{y}_t$. Then, the learner has the option to query any subset of teachers: each teacher charges a unit cost per query and provides a binary label, without any indication of his confidence and without the option of abstaining. The labels received from the queried teachers may disagree, and the learner has no a-priori way of knowing which teacher to trust. If the learner queries the wrong teachers, their labels may agree but still be wrong. The algorithm's goal remains to make accurate predictions using a small number of queries. However, in the absence of a ground truth labeling, it is unclear how to define what it means to make an accurate prediction. To resolve this problem, we formalize

the assumption that different teachers have different areas of expertise, which allows us to compare each predicted label with the labels provided by experts on the relevant topic.

Recalling the motivating example given above, assume that the Internet search company employs multiple human teachers. Some teachers may be better than others across the board and some teachers may be experts on specific topics, such as sports or politics. Some teachers may know the right answer, while others may think they know the right answers but in fact do not—for this reason we do not rely on the teachers themselves to reveal their expertise regions. For example, say that the search engine receives the web query "nhl new york team" and a candidate url is "kings.nhl.com"; a teacher who is a hockey expert would know that this is a bad match (since New York's NHL hockey team is called the Rangers and not the Kings) while a non-expert may not know the answer. The learner has no a-priori knowledge of which teacher to query for the label; yet, in our analysis we would like to compare the learner's prediction to the label given by the expert teacher.

The multiple-teacher selective sampling setting is the focus of the second half of this paper. Specifically, in Section 3 we present a multiple-teacher extension of the (single-teacher) adversarial-stochastic model mentioned earlier, along with two new learning algorithms in this setting. Our model of the teachers' expertise regions enables our algorithms to gradually identify the expertise region of each teacher. Roughly speaking, the algorithm attempts to measure the consistency of the binary labels provided by each teacher in different regions of the instance space. Our first multiple-teacher algorithm has the property that it either queries all of the teachers or does not query any teacher, on each round. Our second algorithm is more sophisticated and queries only those teachers it believes to be experts on $\mathbf{x}_t$. Again, we provide a theoretical analysis that bounds both regret and number of queries issued to the teachers.

Since our results rely on the specific stochastic model of the teachers, it is natural to question how well this model approximates the real-world. To gain some confidence in our assumptions and in our algorithms, in Section 4 we present a simple empirical study on real data that both validate our theoretical results and demonstrates the effectiveness of our approach.

1.1 Related Work in the Single Teacher Setting

Single-teacher selective sampling lies between passive learning (where the algorithm has no control over the learning sequence) and fully active learning (where the learning algorithm is allowed to select the instances $\mathbf{x}_t$). The literature on active learning is vast, and we can hardly do it justice here. Recent papers on active learning include the works by Balcan et al. (2006), Bach (2006), Balcan et al. (2007), Balcan et al. (2008), Castro and Nowak (2008), Dasgupta et al. (2005), Dasgupta et al. (2008), Hanneke (2007), Hanneke (2009) and Koltchinskii (2010). All of these papers consider the case where instances are drawn i.i.d. from a fixed distribution (either known or unknown). In particular, Dasgupta et al. (2005) gives an efficient Perceptron-like algorithm for learning within accuracy ε the class of homogeneous d-dimensional half-spaces under the uniform distribution over the unit ball, with label complexity of the form $d \log \frac{1}{\varepsilon}$. Still in the i.i.d. setting, more general results are given by Balcan et al. (2007). A neat analysis of previously proposed general active learning schemes (Balcan et al., 2006; Dasgupta et al., 2008) is provided by the aforementioned paper by Hanneke (2009). Even more recently, a general Rademacher complexity-based analysis of active learning is given by Koltchinskii (2010). Due to their generality, many of the above results rely on schemes that are computationally prohibitive, exceptions being the results by Dasgupta et al. (2005) and the realizable cases analyzed by Balcan et al. (2007). For instance, the general algorithms

proposed by Hanneke (2009); Koltchinskii (2010) do actually imply estimating ε-minimal sets (or disagreement sets) from empirical data and (local) Rademacher complexities, which makes them computationally hard even for simple function classes, like linear-threshold functions. Finally, pool-based active learning scenarios are considered by Bach (2006) (and the references therein), though the analysis therein is only asymptotic in nature and no quantification is given of the trade-off between risk and number of labels.

To contrast our work with the papers mentioned above, it is worth stressing that our results hold with no stochastic assumption on the source of the instances—in fact, we assume that the instances may be generated by an adaptive adversary. However, as mentioned above, we also show how our online learning algorithm can be converted into a statistical active learning algorithm, with a formal risk bound. Our results in the online selective sampling setting are more in line with the worst-case analyses by Cesa-Bianchi et al. (2006), Strehl and Littman (2008), Cesa-Bianchi et al. (2009) and Orabona and Cesa-Bianchi (2011). These papers present variants of Recursive Least Squares algorithms that operate on arbitrary instance sequences. The analysis by Cesa-Bianchi et al. (2006) is completely worst case: the authors make no assumptions whatsoever on the mechanism generating instances or labels; however, they are unable to prove bounds on the label query rate. The setups by Strehl and Littman (2008), Cesa-Bianchi et al. (2009) and Orabona and Cesa-Bianchi (2011) are closest to ours in that they assume the same stochastic model of the teacher. Our bounds can be shown to be optimal with respect to certain parameters and, unlike competing works on this subject, we are able to face the case when the instance sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ is generated by an *adaptive adversary*, rather than the weaker *oblivious adversary*, as by, for example, Cesa-Bianchi et al. (2009) and Orabona and Cesa-Bianchi (2011). It is actually this difference that makes it possible the selective sampling-to-active learning conversion. A detailed comparison of our results in the single-teacher setting with the results of the predominant papers on this topic is given in Section 2.6, after our results are presented.

1.2 Related Work in the Multiple Teacher Setting

There is also much related work in the multiple-teacher setting, which is often motivated within recent *crowdsourcing* applications. We can map the current state-of-the-art on this topic along various interesting axes.

First, we distinguish between techniques that attempt to find the ground-truth labeling (and evaluate each teacher's quality) independent of the learning algorithm, and techniques that combine the ground-truth-finding and the actual learning into a single algorithm. In the first category are the classical work of Dawid and Skeene (1979), which presents techniques of reconciling conflicting responses on medical questionnaires, the one of Spiegelhalter and Stovin (1983) which handles conflicting information from repeated biopsies, the one by Smyth et al. (1995), where the authors infer a ground truth from multiple annotations of astronomical images, and the one by Hui and Zhou (1998) which examines the more general problem of evaluation in the absence of a ground truth. Still in the first category, Dekel and Shamir (2009a) and Chen et al. (2010) both present general techniques for identifying and rejecting low quality teachers. Papers in the second category discuss supervised learning algorithms that can handle multiple-teacher input. In this category, Dekel and Shamir (2009b) present an SVM variant that is less sensitive to bad labels generated by a small set of malicious teachers, Raykar et al. (2010) use EM to jointly establish a ground truth labeling and learn a maximum-likelihood estimator, Argall et al. (2009) dynamically choose which human

demonstrator to use when teaching a policy to a robot, and Groot et al. (2011) integrate multiple-teacher support into Gaussian process regression learning. Our work in the current paper falls in the second category.

We can also distinguish between algorithms that rely on repeated labeling (where multiple teachers label each example), versus techniques that assume that each example is labeled only once. Sheng et al. (2008), Snow et al. (2008) and Donmez et al. (2009) collect repeated labels and aggregate them (e.g., using a majority vote) to simulate the ground-truth labeling. Some of these papers balance an explore-exploit tradeoff, which determines how many repeated labels are needed for each example. At the opposite end of the spectrum, Dekel and Shamir (2009a) identify low-quality teachers and labels without any repeated labeling. The technique presented in this paper falls in the latter category, since we actively determine which subset of teachers to query on each online round. However, while we do query multiple teachers, we do not assume that the majority vote, or any other aggregate label, is accurate. Still, we do compare to some majority vote of teachers in both our analysis and our experiments.

Next, we distinguish between papers that consider the overall quality score of each teacher (over the entire input space) from papers that assume that each teacher has a specific area of expertise. Most of the papers mentioned above fall in the first category. In the second category, Yan et al. (2010) extend the work in Raykar et al. (2010) (again, maximizing likelihood and using EM) to handle the case where different teachers have knowledge about different parts of the input space. In the present paper, we also model each teacher as an expert on a different subtopic. A closely related research topic is multi-domain adaptation (Mansour et al., 2009a,b), where multiple hypotheses must be optimally combined, under the assumption that each hypothesis makes accurate predictions with respect to a different distribution. Another closely related topic is learning from multiple sources (Crammer et al., 2008), where multiple data sets are sampled from different distributions, and the goal is to optimally combine them with a given target distribution in mind. However, in both of these related problems we are given some prior information on the various distributions, whereas in the multiple-teacher setting we must infer the expertise of each teacher from data.

Another interesting distinction can be made between passive multiple-teacher techniques, which process a static data set that was collected beforehand, and active techniques that route each example to the appropriate teacher. Most of the aforementioned work follows the static approach. The proactive learning setting (Domnez, 2010; Yang and Carbonell, 2009a,b) assumes that the learner has access to teachers of different global quality, with associated costs per label. Yang and Carbonell (2009a) present a theoretical analysis of proactive learning, under the assumption that each teacher gives the correct label most of the time. However, note that the active category fits quite nicely with the assumption that each teacher has an area of expertise (as opposed to measuring the global quality of each teacher): once the algorithm identifies the area of expertise of a teacher, it seems only natural to actively route the relevant examples to that teacher. The approach presented in this paper does precisely that. At the time of writing the extended version of our paper, other works have been published that considered the problem of active learning from multiple annotators. The one whose goal is closest to ours is perhaps the paper by Yan et al. (2011), where a probabilistic multi-labeler model is formulated that allows one to learn the expertise of the labelers and to single out the most uncertain sample (within a given pool of unlabeled instances) whose label is useful to query. Though that paper is similar in spirit to ours, it does mainly focus on modeling and empirical investigations. Finally, we note that Melville et al. (2005) study the closely related problem of actively acquiring individual feature values.

An interesting variation on the multiple-teacher theme involves allowing each teacher's quality to vary with time (Donmez et al., 2010).

2. The Single Teacher Case

In this section, we focus on the standard online selective sampling setting, where the learner has to learn an accurate predictor while determining whether or not to query the label of each instance it observes. We formally define the problem setting in Section 2.1 and introduce our algorithm in Section 2.2. We prove upper bounds on the regret and on the number of queries in Section 2.3. We briefly mention how to convert our online learning algorithm into a statistical active learning algorithm in Section 2.4 and Section 2.5, and we compare our results to related work in Section 2.6.

2.1 Preliminaries and Notation

As mentioned above, on round t of the online selective sampling game, the learner receives an instance $\mathbf{x}_t \in \mathbb{R}^d$, predicts a binary label $\hat{y}_t \in \{-1,+1\}$, and chooses whether or not to query the correct label $y_t \in \{-1,+1\}$. We set $Z_t = 1$ if a query is issued on round t and $Z_t = 0$ otherwise. The only assumption we make on the process that generates $\mathbf{x}_t$ is that $\|\mathbf{x}_t\| \leq 1$; for all we know, instances may be generated by an *adaptive* adversary (an adversary that reacts to our previous actions). Note that most of the previous work on this topic makes stronger assumptions on the process that generates $\mathbf{x}_t$, resulting in a less powerful setting. As for the labels provided by the teacher, we adopt the standard stochastic linear noise model for this problem (Cesa-Bianchi et al., 2003; Cavallanti et al., 2009; Cesa-Bianchi et al., 2009; Strehl and Littman, 2008) and assume that each $y_t \in \{-1,+1\}$ is sampled according to the law

$$P(y_t = 1 \,|\mathbf{x}_t\,) = \frac{1+\mathbf{u}^\top \mathbf{x}_t}{2}, \tag{1}$$

where $\mathbf{u} \in \mathbb{R}^d$ is a fixed but unknown vector with $\|\mathbf{u}\| \leq 1$. Note that $\mathbb{E}\,[y_t \,|\mathbf{x}_t\,] = \mathbf{u}^\top \mathbf{x}_t$, and we denote this value by Δ_t. Unlike much of the recent literature on active learning (see Section 1.1), this simple noise model has the advantage of delivering time-efficient algorithms of practical use.

The learner constructs a sequence of linear predictors $\mathbf{w}_0, \mathbf{w}_1, \ldots$, where each $\mathbf{w}_t \in \mathbb{R}^d$, and predicts $\hat{y}_t = \mathrm{sign}(\hat{\Delta}_t)$ where $\hat{\Delta}_t = {\mathbf{w}_{t-1}}^\top \mathbf{x}_t$. The desirable outcome is for the sequence $\mathbf{w}_0, \mathbf{w}_1, \ldots$ to quickly converge to $\mathbf{u}$. Let P_t denote the conditional probability $\mathbb{P}(\,\cdot\,|\,\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_t, y_1, \ldots, y_{t-1})$. We evaluate the accuracy of the learner's predictions using its *regret*, defined as

$$R_T \;=\; \sum_{t=1}^{T} \big(P_t(y_t \hat{\Delta}_t < 0) - P_t(y_t \Delta_t < 0)\big) \quad .$$

Additionally, we are interested in the number of queries issued by the learner $N_T = \sum_{t=1}^{T} Z_t$. Our goal is to simultaneously bound the regret R_T and the number of queries N_T with high probability over the random draw of labels.

Remark 1 *At first glance, the linear noise model (1) might seem too restrictive. However, this model can be made implicitly nonlinear by running our algorithm in a Reproducing Kernel Hilbert Space $\mathcal{H}$. This entails that the linear operation $\mathbf{u}^\top \mathbf{x}_t$ in (1) is replaced by $h(\mathbf{x}_t)$, for some (typically nonlinear) function $h \in \mathcal{H}$. See also the comments at the end of Section 2.2, and those surrounding Theorem 2.*

2.2 Algorithm

The single teacher algorithm is a margin-based selective sampling procedure. The algorithm "Selective Sampler" (Algorithm 1) depends on a confidence parameter $\delta \in (0,1]$. As in known online ridge-regression-like algorithms (Hoerl and Kennard, 1970; Vovk, 2001; Azoury and Warmuth, 2001; Cesa-Bianchi et al., 2003, 2005a; Li et al., 2008; Strehl and Littman, 2008; Cavallanti et al., 2009; Cesa-Bianchi et al., 2009), our algorithm maintains a weight vector $\mathbf{w}_t$ (initialized as $\mathbf{w}_0 = \mathbf{0}$) and a data correlation matrix A_t (initialized as $A_0 = I$). After receiving $\mathbf{x}_t$ and predicting $\hat{y}_t = \mathrm{sign}(\hat{\Delta}_t)$, the algorithm computes an adaptive data-dependent threshold θ_t, defined as

$$\theta_t^2 \;=\; \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t\left(1+4\sum_{i=1}^{t-1} Z_i r_i + 36\log\frac{t}{\delta}\right) \;,$$

where $r_i = \mathbf{x}_i^\top A_i^{-1}\mathbf{x}_i$. The definition of θ_t follows from our analysis, and can be interpreted as the algorithm's uncertainty in its own predictions. More precisely, the learner believes that $|\hat{\Delta}_t - \Delta_t| \leq \theta_t$. A query is issued only if $|\hat{\Delta}_t| \leq \theta_t$, or in other words, when the algorithm is unsure about the sign of Δ_t. In Algorithm 1, this is denoted by $Z_t = \mathbb{1}\left\{\hat{\Delta}_t^2 \leq \theta_t^2\right\}$, where $\mathbb{1}\left\{\cdot\right\}$ denotes the indicator function.

If the label is not queried, ($Z_t = 0$) then the algorithm does not update its internal state (and $\mathbf{x}_t$ is discarded). If the label is queried ($Z_t = 1$), then the algorithm computes the intermediate vector $\mathbf{w}'_{t-1}$ in such a way that $\hat{\Delta}'_t = {\mathbf{w}'_{t-1}}^\top \mathbf{x}_t$ is at most one in magnitude. Observe that $\hat{\Delta}_t$ and $\hat{\Delta}'_t$ have the same sign and only their magnitudes can differ. In particular, it holds that

$$\hat{\Delta}'_t = \begin{cases} \mathrm{sgn}(\hat{\Delta}_t) & \text{if } |\hat{\Delta}_t| > 1 \\ \hat{\Delta}_t & \text{otherwise} \;. \end{cases}$$

Next, the algorithm defines the new vector $\mathbf{w}_t$ so that $A_t\mathbf{w}_t$ undergoes an additive update, where A_t is a rank-one adjustment of A_{t-1}.

The algorithm can be run both in primal form (as in the pseudocode in Algorithm 1) and in dual form (i.e., in a Reproducing Kernel Hilbert Space). It is not hard to show that the algorithm has a quadratic running time per round, where quadratic means $O(d^2)$ if it is run in primal form, and $O(N_t^2)$ if it is run in dual form, where $N_t = \sum_{i \leq t} Z_i$ is the number of labels requested by the algorithm up to time t. In the dual case, since the algorithm updates only when $Z_t = 1$, the number of labels N_t also corresponds to the number of support vectors used to define the current hypothesis.

2.3 Analysis

Before diving into a formal analysis of Algorithm 1, we attempt to give some intuition regarding our choice of θ_t. Recall that θ_t is the radius of the algorithm's confidence interval, and therefore a small value of θ_t implies that the algorithm is highly confident that Δ_t and $\hat{\Delta}_t$ are close. If, additionally, Δ_t is large, then $\mathrm{sign}(\hat{\Delta}_t)$ is likely to equal $\mathrm{sign}(\Delta_t)$, and the algorithm's prediction is correct. Therefore, we want to show that θ_t can be kept small without issuing an excessive number of queries. To see this, we notice that θ_t depends on the three terms: $\sum_{i=1}^{t-1} Z_i r_i$, $\log(t/\delta)$, and $\mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t$. Later in this section, we prove that $\sum_{i=1}^{t} Z_i r_i$ grows logarithmically with the number of queries N_t, and obviously $\log(t/\delta)$ grows logarithmically with t. To show that θ_t remains small, we must show that the third term, $\mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t$, decreases quickly when labels are queried. $\mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t$ depends on the relationship between the current instance $\mathbf{x}_t$ and the previous instances on rounds where a query was issued.

If $\mathbf{x}_t$ lies along the directions spanned by the previous instances, we show that $\mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t$ tends to shrink as $1/N_t$. As a result, θ_t is on the order of $\log(t/\delta)/N_t$, and N_t only needs to grow at a slow logarithmic rate. On the other hand, if the adversary chooses $\mathbf{x}_t$ outside of the subspace spanned by the previous examples, then the term $\mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t$ causes θ_t to be large, and the algorithm becomes more likely to issue a query. Overall, to ensure a small value of θ_t across the instance space spanned by the $\mathbf{x}_t$ produced by the adversary, the algorithm must query $O\big(\log(t)\big)$ labels in each direction of this instance space.

As noted above, the adversary can arbitrarily inflate our regret by choosing instances that induce small values of Δ_t. Recall that a small value of Δ_t implies that the teacher guesses the label y_t almost at random. Following Cesa-Bianchi et al. (2009), the bounds we prove depend on how many of the instances $\mathbf{x}_t$ are chosen such that Δ_t is very small. Formally, for any $\varepsilon > 0$, define

$$T_\varepsilon = \sum_{t=1}^{T} \mathbf{1}\{|\Delta_t| \le \varepsilon\} \ . \tag{2}$$

The following theorem is the main result of this section, and is stated so as to emphasize both the data-dependent and the time-dependent aspects of our bounds.

Theorem 2 *Assume that Selective Sampler is run with confidence parameter $\delta \in (0,1]$. Then with probability at least $1-\delta$ it holds that for all $T \ge 3$*

$$\begin{aligned}
R_T &\le \inf_{\varepsilon>0}\Big\{\varepsilon T_\varepsilon + \frac{2+8\log|A_T|+144\log(T/\delta)}{\varepsilon}\Big\} = \inf_{\varepsilon>0}\Big\{\varepsilon T_\varepsilon + O\Big(\frac{d\log T+\log(T/\delta)}{\varepsilon}\Big)\Big\} \\
N_T &\le \inf_{\varepsilon>0}\Big\{T_\varepsilon + O\Big(\frac{\log|A_T|\log(T/\delta)+\log^2|A_T|}{\varepsilon^2}\Big)\Big\} = \inf_{\varepsilon>0}\Big\{T_\varepsilon + O\Big(\frac{d^2\log^2(T/\delta)}{\varepsilon^2}\Big)\Big\} \ ,
\end{aligned}$$

where $|A_T|$ is the determinant of the matrix A_T.

Note that the bounds above depend on d the dimension of the instance space. In the case of a (possibly infinite-dimensional) Reproducing Kernel Hilbert Space, d is replaced by a quantity that depends on the spectrum of the data's Gram matrix.

The proof of Theorem 2 splits into a series of lemmas. For every $T > 0$ and $\varepsilon > 0$, we define

$$\begin{aligned}
U_{T,\varepsilon} &= \sum_{t=1}^{T} \bar{Z}_t \ \mathbf{1}\big\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 > \varepsilon^2\big\}, \\
Q_{T,\varepsilon} &= \sum_{t=1}^{T} Z_t \ \mathbf{1}\big\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 > \varepsilon^2\big\}|\Delta_t| \ ,
\end{aligned}$$

where $\bar{Z}_t = 1 - Z_t$. In the above, $U_{T,\varepsilon}$ deals with rounds where the algorithm does not make a query, while $Q_{T,\varepsilon}$ deals with rounds where the algorithm does make a query. The proof exploits the potential-based method for online ridge-regression-like algorithms we learned from Azoury and Warmuth (2001). See also the works of Hazan et al. (2007), Dani et al. (2008) and Crammer and Gentile (2011) for a similar use in different contexts. The potential function we use is the (quadratic) Bregman divergence $d_t(\mathbf{u},\mathbf{w}) = \frac{1}{2}(\mathbf{u}-\mathbf{w})^\top A_t(\mathbf{u}-\mathbf{w})$, where A_t is the matrix computed by Selective Sampler at time t.

The proof structure is as follows. First, Lemma 3 below decomposes the regret R_T into three parts:

$$R_T \le \varepsilon T_\varepsilon + U_{T,\varepsilon} + Q_{T,\varepsilon} \ .$$

Algorithm 1: Selective Sampler

input confidence level $\delta \in (0,1]$
initialize $\mathbf{w}_0 = \mathbf{0},\ A_0 = I$
for $t = 1, 2, \ldots$
  receive $\mathbf{x}_t \in \mathbb{R}^d : ||\mathbf{x}_t|| \leq 1,$ and set $\hat{\Delta}_t = \mathbf{w}_{t-1}{}^\top \mathbf{x}_t$
  predict $\hat{y}_t = \mathrm{sgn}(\hat{\Delta}_t) \in \{-1,+1\}$
  $\theta_t^2 = \mathbf{x}_t^\top A_{t-1}^{-1} \mathbf{x}_t \left(1 + 4\sum_{i=1}^{t-1} Z_i r_i + 36\log(t/\delta)\right)$
  $Z_t = \mathbb{1}\left\{\hat{\Delta}_t^2 \leq \theta_t^2\right\} \in \{0,1\}$
  if $Z_t = 1$
    query $y_t \in \{-1,+1\}$
    $$\mathbf{w}'_{t-1} = \begin{cases} \mathbf{w}_{t-1} - \mathrm{sgn}(\hat{\Delta}_t)\left(\frac{|\hat{\Delta}_t|-1}{\mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t}\right) A_{t-1}^{-1}\mathbf{x}_t & \text{if } |\hat{\Delta}_t| > 1 \\ \mathbf{w}_{t-1} & \text{otherwise} \end{cases}$$
    $A_t = A_{t-1} + \mathbf{x}_t\mathbf{x}_t^\top,\quad r_t = \mathbf{x}_t^\top A_t^{-1}\mathbf{x}_t,\quad \mathbf{w}_t = A_t^{-1}(A_{t-1}\mathbf{w}'_{t-1} + y_t\mathbf{x}_t)$
  else
    $A_t = A_{t-1},\ \mathbf{w}_t = \mathbf{w}_{t-1},\ r_t = 0$

The bound on $U_{T,\varepsilon}$ is given by Lemma 4. For the bound on $Q_{T,\varepsilon}$ and the bound on the number of queries N_T, we use Lemmas 5 and 6, respectively. However, both of these lemmas require that $(\Delta_t - \hat{\Delta}_t)^2 \leq \theta_t^2$ for all t. This assumption is taken care of by the subsequent Lemma 7. Since ε is a positive free parameter, we can take the infimum over $\varepsilon > 0$ to get the required results. In turn, many of these lemmas rely on technical lemmas given in Appendix A and Appendix B.

Lemma 3 *For any $\varepsilon > 0$ it holds that* $R_T \leq \varepsilon T_\varepsilon + U_{T,\varepsilon} + Q_{T,\varepsilon}$.

Proof We have

$$\begin{aligned}
&P_t(\hat{\Delta}_t y_t < 0) - P_t(\Delta_t y_t < 0) \\
&\quad\leq \mathbb{1}\{\hat{\Delta}_t\Delta_t \leq 0\}\left|2P_t(y_t = 1) - 1\right| \\
&\quad= \mathbb{1}\{\hat{\Delta}_t\Delta_t \leq 0\}|\Delta_t| \\
&\quad= \mathbb{1}\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 \leq \varepsilon^2\}|\Delta_t| + \mathbb{1}\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 > \varepsilon^2\}|\Delta_t| \\
&\quad\leq \varepsilon\,\mathbb{1}\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 \leq \varepsilon^2\} + \mathbb{1}\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 > \varepsilon^2\}|\Delta_t| \qquad (3) \\
&\quad= \varepsilon\,\mathbb{1}\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 \leq \varepsilon^2\} + \mathbb{1}\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 > \varepsilon^2, Z_t = 0\}|\Delta_t| \\
&\qquad\qquad + \mathbb{1}\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 > \varepsilon^2, Z_t = 1\}|\Delta_t| \\
&\quad\leq \varepsilon\,\mathbb{1}\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 \leq \varepsilon^2\} + \bar{Z}_t\,\mathbb{1}\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 > \varepsilon^2\} \qquad (4) \\
&\qquad\qquad + Z_t\,\mathbb{1}\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 > \varepsilon^2\}|\Delta_t| \,.
\end{aligned}$$

The inequality in Equation (4) follows directly from $|\Delta_t| \leq 1$. Summing over $t = 1 \ldots T$ completes the proof. ■

Lemma 4 *For any $\varepsilon > 0$ and $T \geq 3$, with probability at least $1-\delta$, it holds that*

$$Q_{T,\varepsilon} \;\leq\; \frac{2+8\log|A_T|+144\log(T/\delta)}{\varepsilon} \;=\; O\left(\frac{d\log T+\log(T/\delta)}{\varepsilon}\right) .$$

Proof We begin with

$$\begin{aligned} Q_{T,\varepsilon} \;&=\; \sum_{t=1}^{T} Z_t \,\mathbb{1}\{\Delta_t\hat{\Delta}_t < 0\}\,\mathbb{1}\{\Delta_t^2 > \varepsilon^2\}|\Delta_t| \\ &\leq\; \frac{1}{\varepsilon}\sum_{t=1}^{T} Z_t \,\mathbb{1}\{\hat{\Delta}_t\Delta_t < 0\}\Delta_t^2 \\ &=\; \frac{1}{\varepsilon}\sum_{t=1}^{T} Z_t \,\mathbb{1}\{\hat{\Delta}'_t\Delta_t < 0\}\Delta_t^2 \;\; . \end{aligned}$$

$\hat{\Delta}'_t\Delta_t < 0$ implies that $\Delta_t^2 \leq (\Delta_t - \hat{\Delta}'_t)^2$, and therefore the above can be upper bounded by

$$\frac{1}{\varepsilon}\sum_{t=1}^{T} Z_t(\Delta_t - \hat{\Delta}'_t)^2.$$

Next we rely on some standard technical results that are given in the appendix. Lemma 23 (i) upper bounds the above by

$$\frac{2}{\varepsilon}\sum_{t=1}^{T} Z_t\,\left((y_t - \hat{\Delta}'_t)^2 - (y_t - \Delta_t)^2\right) + \frac{144}{\varepsilon}\log(T/\delta).$$

Lemma 25 (iv) further bounds this term by

$$\frac{4}{\varepsilon}\sum_{t=1}^{T} Z_t\left(d_{t-1}(\mathbf{u},\mathbf{w}'_{t-1}) - d_t(\mathbf{u},\mathbf{w}'_t) + 2\log\frac{|A_t|}{|A_{t-1}|}\right) + \frac{144}{\varepsilon}\log(T/\delta).$$

After telescoping and using the facts that $d_0(\mathbf{u},\mathbf{w}'_0) = d_0(\mathbf{u},\mathbf{w}_0) = ||\mathbf{u}||^2/2 \leq 1/2$ and $|A_0| = 1$, the above is bounded by

$$\frac{2+8\log|A_T|+144\log(T/\delta)}{\varepsilon},$$

which is in fact $O\left(\frac{d\log T+\log(T/\delta)}{\varepsilon}\right)$ in the finite-dimensional case. This concludes the proof. ■

Lemma 5 *Assume that $(\Delta_t - \hat{\Delta}_t)^2 \leq \theta_t^2$ holds for all t. Then, for any $\varepsilon > 0$, we have $U_{T,\varepsilon} = 0$*

Proof We rewrite our assumption $(\Delta_t - \hat{\Delta}_t)^2 \leq \theta_t^2$ as

$$\Delta_t\hat{\Delta}_t \geq \frac{\hat{\Delta}_t^2 + \Delta^2 - \theta_t^2}{2} \geq \frac{\hat{\Delta}_t^2 - \theta_t^2}{2} \;.$$

However, if $\bar{Z}_t = 1$, then $\hat{\Delta}_t^2 > \theta_t^2$ and so $\Delta_t\hat{\Delta}_t \geq 0$. Hence, under the above assumption, we can guarantee that for any t, $\bar{Z}_t\,\mathbb{1}\{\Delta_t\hat{\Delta}_t < 0\} = 0$, thereby implying $U_{T,\varepsilon} = \sum_{t=1}^{T}\bar{Z}_t\,\mathbb{1}\{\Delta_t\hat{\Delta}_t < 0, \Delta_t^2 > \varepsilon^2\} = 0$. ■

In the proof of the next two lemmas, we use the shorthand $g(t) = \sum_{i=1}^{t} Z_i r_i$.

Lemma 6 *Assume that* $(\Delta_t - \hat{\Delta}_t)^2 \le \theta_t^2$ *holds for all* t*. Then, for any* $\varepsilon > 0$*, and* $T > 0$ *we have*

$$N_T \;\le\; T_\varepsilon + O\left(\frac{\log|A_T|\log(T/\delta) + \log^2|A_T|}{\varepsilon^2}\right) \;=\; T_\varepsilon + O\left(\frac{d^2\log^2(T/\delta)}{\varepsilon^2}\right).$$

Proof Let us rewrite our assumption $(\Delta_t - \hat{\Delta}_t)^2 \le \theta_t^2$ as $|\Delta_t - \hat{\Delta}_t| \le \theta_t$. Then $|\hat{\Delta}_t| \le \theta_t$ implies $|\Delta_t| \le 2\theta_t$. We can write

$$\begin{aligned}
Z_t &= \mathbb{1}\{\hat{\Delta}_t^2 \le \theta_t^2\} \le \mathbb{1}\{\hat{\Delta}_t^2 \le \theta_t^2, \Delta_t^2 \le 4\theta_t^2\} \\
&= \mathbb{1}\left\{\hat{\Delta}_t^2 \le \theta_t^2, \Delta_t^2 \le 4\theta_t^2, \theta_t^2 \ge \frac{\varepsilon^2 \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t}{8\,r_t}\right\} \\
&\quad + \mathbb{1}\left\{\hat{\Delta}_t^2 \le \theta_t^2, \Delta_t^2 \le 4\theta_t^2, \theta_t^2 < \frac{\varepsilon^2 \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t}{8\,r_t}\right\} \\
&\le \mathbb{1}\left\{\hat{\Delta}_t^2 \le \theta_t^2, \theta_t^2 \ge \frac{\varepsilon^2 \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t}{8\,r_t}\right\} + \mathbb{1}\left\{\Delta_t^2 \le 4\theta_t^2, \theta_t^2 < \frac{\varepsilon^2 \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t}{8\,r_t}\right\}. \qquad (5)
\end{aligned}$$

By Lemma 24 (i) we have $\mathbf{x}_t^T A_{t-1}^{-1}\mathbf{x}_t \le 2\,r_t$, hence

$$\mathbb{1}\left\{\Delta_t^2 \le 4\theta_t^2, \theta_t^2 < \frac{\varepsilon^2 \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t}{8\,r_t}\right\} \le \mathbb{1}\{\Delta_t^2 \le \varepsilon^2\}.$$

Plugging back into (5) and summing over t shows that, for any $\varepsilon > 0$,

$$N_T \le T_\varepsilon + \sum_{t=1}^{T} \mathbb{1}\left\{\hat{\Delta}_t^2 \le \theta_t^2, \theta_t^2 \ge \frac{\varepsilon^2 \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t}{8\,r_t}\right\}.$$

Now observe that, by definition of Z_t and θ_t

$$\begin{aligned}
&\sum_{t=1}^{T}\left\{\hat{\Delta}_t^2 \le \theta_t^2, \theta_t^2 \ge \frac{\varepsilon^2 \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t}{8\,r_t}\right\} \\
&\qquad = \sum_{t=1}^{T} Z_t\,\mathbb{1}\left\{8\,r_t\left(1 + 4g(t-1) + 36\log(t/\delta)\right) \ge \varepsilon^2\right\} \\
&\qquad \le \frac{8}{\varepsilon^2}\sum_{t=1}^{T} Z_t\,r_t\left(1 + 4g(t-1) + 36\log(t/\delta)\right).
\end{aligned}$$

Using Lemma 24 (ii), the above is upper bounded by

$$\frac{8}{\varepsilon^2}\,(1 + 36\log(T/\delta))\log|A_T| + \frac{32}{\varepsilon^2}\sum_{t=1}^{T} Z_t r_t g(t-1),$$

which is in turn upper bounded by

$$\frac{8}{\varepsilon^2}\,(1 + 36\log(T/\delta))\log|A_T| + \frac{16}{\varepsilon^2}\sum_{t=1}^{T}\left(g^2(t) - g^2(t-1)\right).$$

Again using Lemma 24 (ii), we upper bound the above by

$$\frac{8}{\varepsilon^2}\,(1+36\log(T/\delta))\log|A_T| + \frac{16}{\varepsilon^2}\,\log^2|A_T|.$$

This term is $O\left(\frac{\log|A_T|\log(T/\delta)+\log^2|A_T|}{\varepsilon^2}\right)$, and specifically $O\left(\frac{d^2\log^2(T/\delta)}{\varepsilon^2}\right)$ in the finite-dimensional case. Since this above holds for any $\varepsilon > 0$, it also holds for the best choice of ε. ■

Lemma 7 *If Selective Sampler is run with confidence parameter $\delta \in (0,1]$, then with probability at least $1-\delta$, the inequality $(\Delta_t - \hat{\Delta}_t)^2 \le \theta_t^2$ holds simultaneously for all $t \ge 3$.*

Proof First note that by Hölder's inequality,

$$(\Delta_t - \hat{\Delta}_t)^2 = ((\mathbf{w}_{t-1} - \mathbf{u})^\top \mathbf{x}_t)^2 \le 2\,\mathbf{x}_t^T A_{t-1}^{-1}\mathbf{x}_t\; d_{t-1}(\mathbf{w}_{t-1},\mathbf{u})\;. \tag{6}$$

Now let $t' := \operatorname{argmax}_{j\le t-1 : Z_j=1} j$, that is, t' is the last round (up to time $t-1$) on which the algorithm issued a query. Then Lemma 25 (i), (ii), (iii), allows us to write

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{t'} Z_i\left((y_i-\hat{\Delta}_i')^2 - (y_i-\Delta_i)^2\right) &\le \sum_{i=1}^{t'-1} Z_i\left(d_{i-1}(\mathbf{u},\mathbf{w}_{i-1}') - d_i(\mathbf{u},\mathbf{w}_i') + 2Z_i r_i\right) \\
&\quad + d_{t'-1}(\mathbf{u},\mathbf{w}_{t'-1}') - d_{t'}(\mathbf{u},\mathbf{w}_{t'}) + 2r_{t'} \\
&\le \frac{1}{2} - d_{t'}(\mathbf{u},\mathbf{w}_{t'}) + 2g(t')\;,
\end{aligned}$$

where the last step comes from the telescoping sum and the fact that

$$d_0(\mathbf{u},\mathbf{w}_0') = d_0(\mathbf{u},\mathbf{w}_0) = \frac{1}{2}\|\mathbf{u}\|^2 \le 1/2.$$

Moreover, by definition of t' we see that $g(t') = g(t-1)$ and $Z_j = 0$ for any $j \in [t'+1, t-1]$. Hence for any such j we have $\mathbf{w}_j = \mathbf{w}_{t'}$. This yields

$$\frac{1}{2}\sum_{i=1}^{t-1} Z_i\left((y_i-\hat{\Delta}_i')^2 - (y_i-\Delta_i)^2\right) \le \frac{1}{2} - d_{t-1}(\mathbf{u},\mathbf{w}_{t-1}) + 2g(t-1)\;.$$

Plugging back into (6) results in

$$(\Delta_t - \hat{\Delta}_t)^2 \le 2\,\mathbf{x}_t^T A_{t-1}^{-1}\mathbf{x}_t\left(1/2 + 2\,g(t-1) - \frac{1}{2}\sum_{i=1}^{t-1} Z_i\left((y_i-\hat{\Delta}_i')^2 - (y_i-\Delta_i)^2\right)\right).$$

A direct application of Lemma 23 (ii) shows that for any given $t \ge 3$, with probability at least $1-\delta/t^2$,

$$(\Delta_t - \hat{\Delta}_t)^2 \le \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t\,(1 + 4\,g(t-1) + 36\log(t/\delta)) = \theta_t^2\;.$$

Finally, a union bound allows us to conclude that $(\Delta_t - \hat{\Delta}_t)^2 \le \theta_t^2$ holds simultaneously for all $t \ge 3$ with probability at least $1-\delta$. ■

Remark 8 *Computing the intermediate vector $\mathbf{w}'_{t-1}$ from $\mathbf{w}_{t-1}$, as defined in Algorithm 1, corresponds to projecting $\mathbf{w}_{t-1}$ onto the convex set $C_t = \{\mathbf{w} \in \mathbb{R}^d : |\mathbf{w}^\top \mathbf{x}_t| \leq 1\}$ w.r.t. the Bregman divergence d_{t-1}, that is, $\mathbf{w}'_{t-1} = \operatorname{argmin}_{\mathbf{u}\in C_t} d_{t-1}(\mathbf{u}, \mathbf{w}_{t-1})$. Notice that C_t includes the unit ball since $\mathbf{x}_t$ is normalized. This projection step is needed for technical purposes during the construction of a suitable bounded-variance martingale difference sequence (see Lemma 23 in Appendix A). Unlike similar constructions (Hazan et al., 2007; Dani et al., 2008), we do not project onto the unit ball. In fact, computing the latter would involve a line search over matrices, which would significantly slow down the algorithm. On the other hand, it is also interesting to observe that Selective Sampler performs the projection onto C_t only a logarithmic number of times. This is because*

$$\begin{aligned}\sum_{t=1}^{T} \mathbf{1}\{\hat{\Delta}_t^2 \leq \theta_t^2, |\hat{\Delta}_t| > 1\} &\leq \sum_{t=1}^{T} Z_t \hat{\Delta}_t^2 \\ &\leq \sum_{t=1}^{T} Z_t \theta_t^2 \\ &\leq 2\sum_{t=1}^{T} Z_t\, r_t\Big(1 + 4\,g(t-1) + 36\log(t/\delta)\Big),\end{aligned}$$

which is $O\left(d^2 \log^2(T/\delta)\right)$ by Lemma 24 (iii).

2.4 An Online-to-Batch Conversion

It is instructive to see what the bound in Theorem 2 looks like when we assume that the instances $\mathbf{x}_t$ are drawn i.i.d. according to an unknown distribution over the Euclidean unit sphere, and to compare this bound to standard statistical learning bounds. We model the distribution of the instances near the hyperplane $\{\mathbf{x} : \mathbf{u}^\top \mathbf{x} = 0\}$ using the well-known *Mammen-Tsybakov low noise condition* (Tsybakov, 2004):[2]

$$\text{There exist } c > 0 \text{ and } \alpha \geq 0 \text{ such that } P\big(|\mathbf{u}^\top \mathbf{x}| < \varepsilon\big) \leq c\,\varepsilon^\alpha \text{ for all } \varepsilon > 0.$$

We now describe a simple randomized algorithm which, with high probability over the sampling of the data, returns a linear predictor with a small expected risk (expectation is taken over the randomization of the algorithm). The algorithm is as follows:

1. Run Algorithm 1 with confidence level δ on the data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$, and obtain the sequence of predictors $\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_{T-1}$

2. Pick $r \in \{0, 1, \ldots, T-1\}$ uniformly at random and return $\mathbf{w}_r$.

Due to the unavailability of all labels, standard conversion techniques that return a single deterministic hypothesis (Cesa-Bianchi and Gentile, 2008) do not readily apply here. The following theorem, whose proof is given in Appendix C, states a high probability bound on the risk and the label complexity of our algorithm.

2. The constant c might actually depend on the input dimension d. For notational simplicity, Theorem 9 regards c as a constant, hence it is hidden in the big-oh notation.

Theorem 9 *Let $\mathbf{w}_r$ be the linear hypothesis returned by the above algorithm. Then with probability at least $1-\delta$ we have*

$$\mathbb{E}_r\left[P'_r(y\,\mathbf{w}_r^\top\mathbf{x}<0)\right] \le P(y\,\mathbf{u}^\top\mathbf{x}<0) + O\left(\frac{(d\log(T/\delta))^{\frac{\alpha+1}{\alpha+2}}}{T^{\frac{\alpha+1}{\alpha+2}}} + \frac{\log\left(\frac{\log T}{\delta}\right)}{T}\right) \ ,$$

$$N_T = O\left((d^2\log^2(T/\delta))^{\frac{\alpha}{\alpha+2}}\,T^{\frac{2}{\alpha+2}} + \log(1/\delta)\right) \ ,$$

where $\mathbb{E}_r$ is the expectation over the randomization in the algorithm, and $P'_r(\cdot)$ denotes the conditional probability $P(\cdot\,|\,\mathbf{x}_1,\ldots,\mathbf{x}_{r-1},y_1,\ldots,y_{r-1})$.[3]

As α goes from 0 (no assumptions on the noise) to ∞ (hard separation assumption), the above bound on the average regret roughly interpolates between $1/\sqrt{T}$ and $1/T$. Correspondingly, the bound on the number of labels N_T goes from T to $\log^2 T$. In particular, observe that, viewed as a function of N_T (and disregarding log factors), the instantaneous regret is of the form $N_T^{-\frac{\alpha+1}{2}}$. These bounds are sharper than those by Cavallanti et al. (2009) and, in fact, no further improvement is generally possible (Castro and Nowak, 2008). The same rates are obtained by Hanneke (2009) under much more general conditions, for less efficient algorithms that are based on empirical risk minimization.

2.5 Statistical Active Learning

We now briefly show how to turn our algorithm into a standard statistical active learning algorithm.

Following Koltchinskii (2010), we consider a sequential learning protocol for active learning where on round t the algorithm has to choose a subset S_t of the instance space $\{\mathbf{x}\in R^d : ||\mathbf{x}|| \le 1\}$ (the Euclidean sphere) from which the next instance $\mathbf{x}_t$ is sampled from. Specifically, $\mathbf{x}_t$ is sampled from the conditional distribution $P(\cdot\,|\mathbf{x}\in S_t)$, being $P(\cdot)$ an unknown distribution over the Euclidean sphere. The algorithm then observes the associated label y_t, generated according to the linear noise model of Section 2.1. Notice that the set S_t is typically depending on past examples $(\mathbf{x}_1,y_1),\ldots(\mathbf{x}_{t-1},y_{t-1})$. Again, the goal is to study the high probability behavior of the regret as a function of the number of observed labels (which now coincides with the number of sampled instances $\mathbf{x}_t$).

The analysis developed in Section 2.4 immediately accommodates this model of learning, once we let S_t be the querying region of Algorithm 1, that is,

$$S_t = \{\mathbf{x} : (\mathbf{w}_{t-1}^\top\mathbf{x})^2 \le \theta_t^2\},$$

and think of the randomized i.i.d. algorithm described in that section as operating as follows. We sample an independent new instance $\mathbf{x}$ from the Euclidean sphere, and check whether $\mathbf{x}\in S_t$ or not. In the former case, the associated label y is sampled, and the subset S_t is updated into S_{t+1} according to the rules of Algorithm 1 for updating $\mathbf{w}_{t-1}$ into $\mathbf{w}_t$ and θ_t into θ_{t+1}. In the latter case, $\mathbf{x}$ is discarded, S_t remains unchanged, and a new independent instance is drawn. Notice that this is precisely what Algorithm 1 does when running with an i.i.d. sequence of examples. The same conclusions we have drawn from Theorem 9 can be repeated here.

3. Notice the difference with the conditional probability $P_r(\cdot)$ defined in Section 2.1.

2.6 Related Work

As we mentioned in Section 1.1, the results of Theorem 2 are more in line with the worst-case analyses by Strehl and Littman (2008), Cesa-Bianchi et al. (2009) and Orabona and Cesa-Bianchi (2011). These papers present variants of Recursive Least Squares algorithms that operate on arbitrary instance sequences, but assuming the same linear stochastic noise-model used in our analysis. The algorithm presented by Strehl and Littman (2008) approximates the Bayes margin to within a given accuracy ε, and queries $\tilde{O}(d^3/\varepsilon^4)$ labels; this bound is significantly inferior to our bound, and it seems to hold only in the finite-dimensional case. A more precise comparison can be made to the (expectation) bounds presented by Cesa-Bianchi et al. (2009) and Orabona and Cesa-Bianchi (2011), which are of the form $R_T \leq \min_{0<\varepsilon<1}\left(\varepsilon T_\varepsilon + \frac{T^{1-\kappa}}{\varepsilon} + \frac{d}{\varepsilon^2}\ln T\right)$, and $N_T = O(dT^\kappa \ln T)$, where $\kappa \in [0,1]$ is a tunable parameter of their algorithm. After a proper setting of κ, this gives rise to an instantaneous regret which is still (up to log factors) in the form $N_T^{-\frac{\alpha+1}{2}}$ under the same low-noise assumptions as in Section 2.4. On the other hand, our bound here does not require tuning of parameters. More importantly, whereas the analysis of Cesa-Bianchi et al. (2009) and Orabona and Cesa-Bianchi (2011) only holds for oblivious adversaries, we cover the case where the instances can be generated adaptively.[4] We emphasize that it is just the adaptivity of the adversary that enabled us to convert our selective sampling algorithm to the statistical active learning algorithm presented in Section 2.5.

Another relevant line of research that came to our attention at the time of writing the extended version of our paper is the importance sampling-based active learning schemes followed by Beygelzimer et al. (2010, 2011). These papers are interesting in that they give up with the version space approach followed by their predecessors (Dasgupta et al., 2008; Hanneke, 2007, 2009; Koltchinskii, 2010) which might deliver time-efficient active learning schemes. A direct comparison to Beygelzimer et al. (2010, 2011) is not straightforward. While we can see that their label selection mechanism (e.g., Algorithm 1 in Beygelzimer et al., 2010) gets similar to the one in our Selective Sampler (once it is adapted to square loss and the class conditional distribution is (1)), their analysis (e.g., Theorem 4 therein) seems to provide suboptimal results. For instance, under hard separation assumptions, their bound on N_T never gets as small as a logarithmic function in T. In short, we suspect that their algorithm (or variants thereof) is a strict generalization of ours but, being more general, the associated analysis is also significantly looser.

3. The Multiple Teacher Case

The problem is still online binary classification, where on each round $t = 1, 2, \ldots$ the learner receives an input $\mathbf{x}_t \in \mathbb{R}^d$, with $\|\mathbf{x}_t\| \leq 1$, and outputs a binary prediction $\hat{y}_t$. However, there are now K available teachers, each with his own area of expertise. The expertise area of each teacher is unknown to the algorithm, and can only be inferred indirectly from the binary labels provided by that teacher and by other teachers. If $\mathbf{x}_t$ falls within the expertise region of teacher j, then that teacher can provide an accurate label. After making each binary prediction, the learner chooses if to issue a query to one or more of the K teachers. The learner is free to query any subset of teachers, but each

4. It is fair to say that Orabona and Cesa-Bianchi (2011) have further improvements over both Cesa-Bianchi et al. (2009) and this paper. In particular, the DGS-Mod algorithm therein is able to handle the case when the vector $\mathbf{u}$ generating the labels has unknown length $||\mathbf{u}||$. However, it does so at the cost of an exponential dependence of R_T on $||\mathbf{u}||$.

teacher charges a unit cost per label. We emphasize that a queried teacher provides only a binary label, and does not indicate his level of confidence in that label.

From the point of view of our learning algorithm, these confidence levels have to be interpreted as *reliability* rates of the teachers. Since these rates play a major role in weighting the relative importance of the teachers, it looks wiser to let the algorithm compute these rates as a function of past interactions among teachers, rather than relying on human "self-judgement".

Formally, we assume that teacher j is associated with a weight vector $\mathbf{u}_j \in \mathbb{R}^d$, where $\|\mathbf{u}_j\| \leq 1$. If teacher j is queried on round t, he stochastically generates the binary label $y_{j,t}$ according to $P_t(y_{j,t} = 1|\mathbf{x}_t) = (1+\Delta_{j,t})/2$, where $\Delta_{j,t} = \mathbf{u}_j{}^\top \mathbf{x}_t$ and, as in Section 2, $\mathbf{x}_t$ can be chosen adversarially depending on previous $\mathbf{x}$'s and y_j's. We consider $|\Delta_{j,t}|$ to be the (hidden) *confidence* of teacher j in his label for $\mathbf{x}_t$. When the learner issues a query, he receives nothing other than the binary label itself, and the confidence is only part of our theoretical model of the teacher. If $\mathbf{x}_t$ is almost orthogonal to $\mathbf{u}_j$ then teacher j has a very low confidence in his label, and we say that $\mathbf{x}_t$ lies outside the expertise region of teacher j.

It is no longer clear how we should evaluate the performance of the learner, since the K teachers will often give inconsistent labels on the given $\mathbf{x}_t$, and we do not have a well-defined ground-truth to compare against. Intuitively, we would like the learner to predict the label of $\mathbf{x}_t$ as accurately as the teachers who are experts on $\mathbf{x}_t$. To formalize this intuition,[5] define the average margin of a generic subset of teachers $C \subseteq [K]$ as $\Delta_{C,t} = \frac{1}{|C|}\sum_{i \in C} \Delta_{i,t}$. We define the set of experts for each instance using a user-specified parameter $\tau > 0$. Define

$$j_t^\star \;=\; \text{argmax}_j |\Delta_{j,t}| \quad \text{and} \quad C_t \;=\; \{i : |\Delta_{i,t}| \geq |\Delta_{j_t^\star,t}| - \tau\} \;. \tag{7}$$

In words, $j_t^\star$ is the *most confident teacher* at time t, and C_t is the *set of confident teachers* at time t. Again, recall that C_t is unknown to the learning algorithm. In this setting, τ is a tolerance parameter that defines how confident a teacher must be, compared to the most confident teacher, to be considered a confident teacher. Although τ does not appear explicitly in the notation C_t, the reader should keep in mind that C_t and other sets defined later on in this section all depend on τ. Using the definitions above, $\Delta_{C_t,t}$ is the average margin of the confident teachers, and we abbreviate $\Delta_t = \Delta_{C_t,t}$.

Now, let y_t be the random variable that takes values in $\{-1,1\}$, with $P_t(y_t = 1|\mathbf{x}_t) = (1+\Delta_t)/2$. In words, y_t is the binary label generated according to the average margin of the confident teachers. We consider the sequence $y_1, \ldots, y_T$ to be our ad-hoc ground-truth, and the goal of our algorithm is to accurately predict this sequence. Note that an equivalent way of generating y_t is to pick a confident teacher j uniformly at random from C_t and to set $y_t = y_{j,t}$. Indeed there are other reasonable ways to define the ground-truth for this problem, however, we feel that our definition coincides with our intuitions on learning from teachers with different areas of expertise. If τ is set to 1, the learner is compared against the average margin of all K teachers, while if $\tau = 0$, the learner is compared against the single most confident teacher.

Remark 10 *The reader might wonder whether the framework just described could be accommodated by a standard experts setting (e.g., Cesa-Bianchi and Lugosi, 2006) or, perhaps, by a label-efficient version thereof (e.g., Helmbold and Panizza, 1997; Cesa-Bianchi et al., 2005b). Due to the absence of a ground truth, the answer is negative. Of course, we might be tempted to apply*

5. Here and throughout, $[K] = \{1, 2, \ldots, K\}$.

a label-efficient expert algorithm by pretending that the missing ground-truth is provided by some function of the teachers we query. Unfortunately, the above references contain results which are too general to yield tight bounds for our specific noise model. Indeed, our ambition here is to leverage the side information provided by the instance vectors so as to outperform the best single expert in hindsight while, at the same time, querying just a small fraction of the available teachers' labels.

We now describe and analyze two algorithms within the multiple teacher setting. We call these algorithms "first version" and "second version". In the first version, the algorithm queries either all of the teachers or none of them. The second version is more refined in that the algorithm may query a different subset of teachers on each round. In Section 4 we present experiments on real-world data with the second version of the algorithm.

3.1 Algorithm, First Version

The learner attempts to model each weight vector $\mathbf{u}_j$ with a corresponding weight vector $\mathbf{w}_{j,t}$. As in the single teacher case, the learner maintains a variable threshold θ_t, which can be interpreted as the learner's confidence in its current set of weight vectors. The learner attempts to mimic the process of generating y_t by choosing its own set of confident teachers on each round. Denoting $\hat{\Delta}_{j,t} = \mathbf{w}_{j,t}{}^{\top}\mathbf{x}_t$, the learner defines

$$\hat{j}_t \;=\; \text{argmax}_j|\hat{\Delta}_{j,t}| \quad \text{and} \quad \hat{C}_t \;=\; \{i : |\hat{\Delta}_{i,t}| \geq |\hat{\Delta}_{\hat{j}_t,t}| - \tau - 2\theta_t\} \;\;,$$

where $\hat{j}_t$ is the learner's estimate of the most confident teacher, and $\hat{C}_t$ is the learner's estimate of the set of confident teachers. Note that the definition of $\hat{C}_t$ is more inclusive than the definition of C_t in Equation (7), in that it also includes teachers whose confidence falls below $|\hat{\Delta}_{\hat{j}_t,t}| - \tau$. This accounts for the uncertainty regarding the learner's set of weight vectors.

As above, we define the notation $\hat{\Delta}_{C,t} = \frac{1}{|C|}\sum_{i\in C}\hat{\Delta}_{i,t}$, and abbreviate $\hat{\Delta}_t = \hat{\Delta}_{\hat{C}_t,t}$. The learner predicts the binary label $\hat{y}_t = \text{sgn}(\hat{\Delta}_t)$. Let P_t denote the conditional probability $P_t(\cdot) = \mathbb{P}(\cdot|\mathbf{x}_1, y_{1,1}\ldots, y_{K,1}, \mathbf{x}_2, y_{1,2}\ldots, y_{K,2}, \ldots \mathbf{x}_{t-1}, y_{1,t-1}, \ldots y_{K,t-1}, \mathbf{x}_t)$, and define the regret of the learner as

$$R_T \;=\; \sum_{t=1}^{T}\left(P_t(y_t\hat{\Delta}_t < 0) - P_t(y_t\Delta_t < 0)\right) \;\;. \tag{8}$$

Next, we proceed to describe our criterion for querying teachers. We present a simple criterion that either sets $Z_t = 1$ and queries all of the teachers or sets $Z_t = 0$ and queries none of them. Therefore, the learner either incurs a cost of K or a cost of 0 on each round. We partition the set of confident teachers $\hat{C}_t$ into two sets,

$$\begin{aligned}\hat{H}_t \;&=\; \{i \;:\; |\hat{\Delta}_{i,t}| \;\geq\; |\hat{\Delta}_{\hat{j}_t,t}| - \tau + 2\theta_t\},\\ \hat{B}_t \;&=\; \{i \;:\; |\hat{\Delta}_{\hat{j}_t,t}| - \tau - 2\theta_t \;\leq\; |\hat{\Delta}_{i,t}| \;<\; |\hat{\Delta}_{\hat{j}_t,t}| - \tau + 2\theta_t\}\;.\end{aligned}$$

In words, $\hat{H}_t$ is the set of teachers with especially high confidence, while $\hat{B}_t$ is the set of teachers with borderline confidence. Intuitively, the learner is unsure whether the teachers in $\hat{B}_t$ should or should not be included in $\hat{C}_t$. The learner issues a query (to all K teachers) in one of two cases. The first case is when there exists a subset of borderline teachers $S \subseteq \hat{B}_t$ that causes the predicted label to flip, namely, $\hat{\Delta}_t\hat{\Delta}_{\hat{H}_t\cup S,t} < 0$. The second case is when there exists a subset of borderline teachers

Algorithm 2: Multiple Teacher Selective Sampler—first version

input confidence level $\delta \in (0,1]$, tolerance parameter $\tau \geq 0$

initialize $A_0 = I$, $\forall j \in [K]$ $\mathbf{w}_{j,0} = \mathbf{0}$

for $t = 1, 2, \ldots$

receive $\mathbf{x}_t \in \mathbb{R}^d : ||\mathbf{x}_t|| \leq 1$

$\theta_t^2 = \mathbf{x}_t^\top A_{t-1}^{-1} \mathbf{x}_t \left(1 + 4\sum_{i=1}^{t-1} Z_i r_i + 36\log(Kt/\delta)\right)$

$\forall j \in [K]$ $\hat{\Delta}_{j,t} = \mathbf{w}_{j,t-1}{}^\top \mathbf{x}_t$ and $\hat{j}_t = \operatorname{argmax}_j |\hat{\Delta}_{j,t}|$

predict $\hat{y}_t = \operatorname{sgn}(\hat{\Delta}_t) \in \{-1, +1\}$

$$Z_t = \begin{cases} 1 & \text{if } \exists S \subseteq \hat{B}_t : \hat{\Delta}_t \hat{\Delta}_{S \cup \hat{H}_t, t} < 0 \text{ or } |\hat{\Delta}_{S \cup \hat{H}_t, t}| \leq \theta_t \\ 0 & \text{otherwise} \end{cases}$$

if $Z_t = 1$

query $y_{1,t}, \ldots, y_{K,t}$

$A_t = A_{t-1} + \mathbf{x}_t \mathbf{x}_t^\top$, $r_t = \mathbf{x}_t^\top A_t^{-1} \mathbf{x}_t$

for $j = 1, \ldots, K$

$$\mathbf{w}'_{j,t-1} = \begin{cases} \mathbf{w}_{j,t-1} - \operatorname{sgn}(\hat{\Delta}_{j,t}) \left(\frac{|\hat{\Delta}_{j,t}| - 1}{\mathbf{x}_t^\top A_{t-1}^{-1} \mathbf{x}_t} \right) A_{t-1}^{-1} \mathbf{x}_t & \text{if } |\hat{\Delta}_{j,t}| > 1, \\ \mathbf{w}_{j,t-1} & \text{otherwise} \end{cases}$$

$$\mathbf{w}_{j,t} = A_t^{-1}(A_{t-1}\mathbf{w}'_{j,t-1} + y_{j,t}\mathbf{x}_t)$$

else

$A_t = A_{t-1}$, $r_t = 0$ and $\mathbf{w}_{j,t} = \mathbf{w}_{j,t-1}$ $\forall j \in [K]$

$S \subseteq \hat{B}_t$ that causes the margin to be too small, namely $|\hat{\Delta}_{\hat{H}_t \cup S, t}| \leq \theta_t$. In either of these cases, we say that the set of (estimated) confident teachers is *unstable*. If a query is issued, each weight vector $\mathbf{w}_{j,t}$ is updated as in the single teacher case. The pseudocode of this algorithm is given in Algorithm 2.

Remark 11 *At first sight, it may seem that computing Z_t causes an exponential explosion due to the need to check all possible subsets $S \subseteq \hat{B}_t$. The same implementation issue arises in Algorithm 3 (Section 3.3). As a matter of fact, this check can be computed efficiently by first sorting the teachers according to their estimated confidence $|\hat{\Delta}_{j,t}|$, and then greedily growing the subset S by following this order.*

3.2 Analysis, First Version

Our learning algorithm relies on labels it receives from a set of teachers, and therefore our bounds should naturally depend on the ability of those teachers to provide accurate labels for the sequence $\mathbf{x}_1, \ldots, \mathbf{x}_T$. For example, if an input $\mathbf{x}_t$ lies outside the expertise regions of all teachers, we cannot hope to learn anything from the labels provided by the teachers for this input. Similarly, there

is nothing we can do on rounds where the set of confident teachers is split between two equally confident but conflicting opinions. We count these difficult rounds by defining, for any $\varepsilon > 0$,

$$T_\varepsilon \;=\; \sum_{t=1}^{T} \mathbf{1}\{|\Delta_t| \le \varepsilon\} \;. \tag{9}$$

The above is just a multiple teacher counterpart to (2). However it is interesting to note that even in a case where most teachers have low confidence in their prediction on any given round, T_ε can still be small provided that the experts in the field have a confident opinion.

A more subtle difficulty presents itself when the collective opinion expressed by the set of confident teachers changes qualitatively with a small perturbation of the input $\mathbf{x}_t$ or one of the weight vectors $\mathbf{u}_j$. To state this formally, define for any $\varepsilon > 0$

$$\begin{aligned} H_{\varepsilon,t} \;&=\; \{i \;:\; |\Delta_{i,t}| \;\ge\; |\Delta_{j_t^\star,t}| - \tau + \varepsilon\}, \\ B_{\varepsilon,t} \;&=\; \{i \;:\; |\Delta_{j_t^\star,t}| - \tau - \varepsilon \;\le\; |\Delta_{i,t}| \;<\; |\Delta_{j_t^\star,t}| - \tau + \varepsilon\} \;. \end{aligned}$$

The set $H_{\varepsilon,t}$ is the subset of teachers in C_t with especially high confidence, ε higher than the minimal confidence required for inclusion in C_t. In contrast, the set $B_{\varepsilon,t}$ is the set of teachers with borderline confidence: either teachers in C_t that would be excluded if their margin were smaller by ε, or teachers that are not in C_t that would be included if their margin were larger by ε. We say that the average margin of the confident teachers is *unstable* with respect to τ and ε if $|\Delta_t| > \varepsilon$ but we can find a subset $S \subseteq B_{\varepsilon,t}$ such that either $\Delta_t \Delta_{S \cup H_{\varepsilon,t},t} < 0$ or $|\Delta_{S \cup H_{\varepsilon,t},t}| < \varepsilon$. In other words, we are dealing with the situation where Δ_t is sufficiently confident, but a small ε-perturbation to the margins of the individual teachers can cause its sign to flip, or its confidence to fall below ε. We count the unstable rounds by defining, for any $\varepsilon > 0$,[6]

$$T'_\varepsilon \;=\; \sum_{t=1}^{T} \mathbf{1}\big\{|\Delta_t| > \varepsilon\big\}\, \mathbf{1}\big\{\exists S \subseteq B_{\varepsilon,t} \;:\; \Delta_t \Delta_{S \cup H_{\varepsilon,t},t} < 0 \;\vee\; |\Delta_{S \cup H_{\varepsilon,t},t}| \le \varepsilon\big\} \;. \tag{10}$$

Intuitively T'_ε counts the number of rounds on which an ε-perturbation of $\Delta_{t,j}$ either changes the sign of the average margin or results in an average margin close to zero. Like T_ε, this quantity measures an inherent hardness of the multiple teacher problem.

The following theorem is the main theoretical result of this section. It provides an upper bound on the regret of the learner, as defined in Equation (8), and on the total cost of queries, $N_T = K \sum_{t=1}^{T} Z_t$. Again, we emphasize both the data and the time-dependent aspects of the bound.

6. Notice that, up to degenerate cases, both T_ε and T'_ε tend to vanish as $\varepsilon \to 0$. Hence, as in the single teacher case, the free parameter ε trades-off hardness terms against large deviation terms.

Theorem 12 *Assume Algorithm 2 is run with a confidence parameter $\delta > 0$. Then with probability at least $1-\delta$ it holds for all $T \geq 3$ that*

$$\begin{aligned}
R_T &\leq \inf_{\varepsilon>0}\left\{\varepsilon T_\varepsilon + T'_\varepsilon + O\left(\frac{\log|A_T|\,\log(KT/\delta)+\log^2|A_T|}{\varepsilon^2}\right)\right\} \\
&= \inf_{\varepsilon>0}\left\{\varepsilon T_\varepsilon + T'_\varepsilon + O\left(\frac{d^2\log^2(KT/\delta)}{\varepsilon^2}\right)\right\} \ , \\
N_T &\leq K \inf_{\varepsilon>0}\left\{T_\varepsilon + T'_\varepsilon + O\left(\frac{\log|A_T|\,\log(KT/\delta)+\log^2|A_T|}{\varepsilon^2}\right)\right\} \\
&= K \inf_{\varepsilon>0}\left\{T_\varepsilon + T'_\varepsilon + O\left(\frac{d^2\log^2(KT/\delta)}{\varepsilon^2}\right)\right\} \ .
\end{aligned}$$

As in the proof of Theorem 2, we begin by decomposing the regret and the number of queries. Recall the definitions of T_ε and T'_ε in Equation (9) and Equation (10), respectively. Additionally, define for any $\varepsilon > 0$

$$\begin{aligned}
U_T &= \textstyle\sum_{t=1}^T \bar{Z}_t \, \mathbb{1}\{\Delta_t \hat{\Delta}_t < 0\}, \\
Q_{T,\varepsilon} &= \textstyle\sum_{t=1}^T Z_t \, \mathbb{1}\{\forall S \subseteq B_{\varepsilon,t} \ : \ \Delta_t \Delta_{S \cup H_{\varepsilon,t}} \geq 0, \ |\Delta_{S \cup H_{\varepsilon,t}}| > \varepsilon\} \ .
\end{aligned} \tag{11}$$

T_ε and T'_ε deal with rounds on which the ground truth itself is unreliable, U_T sums over rounds where the learner does not issue a query, and $Q_{T,\varepsilon}$ sums over rounds where the learner does issue a query. Using these definitions, we state the following decomposition lemmas.

Lemma 13 *For any $\varepsilon > 0$ it holds that $R_T \ \leq \ \varepsilon T_\varepsilon + T'_\varepsilon + U_T + Q_{T,\varepsilon}$.*

Lemma 14 *For any $\varepsilon > 0$, it holds that $N_T \leq K\,(T_\varepsilon + T'_\varepsilon + Q_{T,\varepsilon})$.*

The proofs of these lemmas are given in Appendix C. To conclude the proof of Theorem 12, it remains to upper bound U_T and $Q_{T,\varepsilon}$.

Lemma 15 *If $(\Delta_{j,t} - \hat{\Delta}_{j,t})^2 \leq \theta_t^2$ holds for all $j \in [K]$ and $t \in [T]$, then*

$$Q_{T,\varepsilon} = O\left(\frac{\log|A_T|\,\log(KT/\delta)+\log^2|A_T|}{\varepsilon^2}\right) = O\left(\frac{d^2\log^2(KT/\delta)}{\varepsilon^2}\right) \ .$$

Lemma 16 *If $(\Delta_{j,t} - \hat{\Delta}_{j,t})^2 \leq \theta_t^2$ for all $j \in [K]$ and $t \in [T]$, then $U_T = 0$.*

The proofs of these lemmas are also given in Appendix C. Both lemmas rely on the assumption that $(\Delta_{j,t} - \hat{\Delta}_{j,t})^2 \leq \theta_t^2$ for all $t \in [T]$ and $j \in [K]$. A straightforward stratification of Lemma 7 in Section 2 over the K teachers verifies that this condition holds with high probability. This concludes the proof of Theorem 2.

Algorithm 3: Multiple Teacher Selective Sampler—second version

input confidence level $\delta \in (0,1]$, tolerance parameter $\tau \geq 0$

initialize $A_{j,0} = I$, $\mathbf{w}_{j,0} = \mathbf{0}$, $\forall j \in [K]$

for $t = 1, 2, \ldots$

receive $\mathbf{x}_t \in \mathbb{R}^d$: $||\mathbf{x}_t|| \leq 1$

$\forall j \in [K]$, $\theta^2_{j,t} = \mathbf{x}_t^\top A^{-1}_{j,t-1} \mathbf{x}_t \left(1 + 4\sum_{i=1}^{t-1} Z_i r_{j,i} + 36\log(Kt/\delta)\right)$

$\forall j \in [K]$, $\hat{\Delta}_{j,t} = \mathbf{w}_{j,t-1}{}^\top \mathbf{x}_t$ and $\hat{j}_t = \text{argmax}_j |\hat{\Delta}_{j,t}|$

predict $\hat{y}_t = \text{sgn}(\hat{\Delta}_t) \in \{-1, +1\}$

$$Z_t = \begin{cases} 1 & \text{if } \exists S \subseteq \hat{B}_t \,:\, \hat{\Delta}_t \hat{\Delta}_{S \cup \hat{H}_t, t} < 0 \quad \text{or} \quad |\hat{\Delta}_{S \cup \hat{H}_t, t}| \leq \theta_{S \cup \hat{H}_t, t} \\ 0 & \text{otherwise} \end{cases}$$

if $Z_t = 1$ and $j \in \hat{C}_t$

query $y_{j,t}$

$A_{j,t} = A_{j,t-1} + \mathbf{x}_t \mathbf{x}_t^\top$, $r_{j,t} = \mathbf{x}_t^\top A^{-1}_{j,t} \mathbf{x}_t$

$$\mathbf{w}'_{j,t-1} = \begin{cases} \mathbf{w}_{j,t-1} - \text{sgn}(\hat{\Delta}_{j,t}) \left(\frac{|\hat{\Delta}_{j,t}| - 1}{\mathbf{x}_t^\top A^{-1}_{j,t-1} \mathbf{x}_t} \right) A^{-1}_{j,t-1} \mathbf{x}_t & \text{if } |\hat{\Delta}_{j,t}| > 1, \\ \mathbf{w}_{j,t-1} & \text{otherwise} \end{cases}$$

$\mathbf{w}_{j,t} = A^{-1}_{j,t} (A_{j,t-1} \mathbf{w}'_{j,t-1} + y_{j,t} \mathbf{x}_t)$

else

$A_{j,t} = A_{j,t-1}$, $r_{j,t} = 0$ and $\mathbf{w}_{j,t} = \mathbf{w}_{j,t-1}$

3.3 Algorithm, Second Version

The second version differs from the first one in that now each teacher j has its own threshold $\theta_{j,t}$, and also its own matrix $A_{j,t}$. As a consequence, the set of confident teachers $\hat{C}_t$ and the partition of $\hat{C}_t$ into highly confident ($\hat{H}_t$) and borderline ($\hat{B}_t$) teachers have to be redefined as follows:

$$\begin{aligned} \hat{C}_t &= \{j : |\hat{\Delta}_{j,t}| \geq |\hat{\Delta}_{\hat{j}_t,t}| - \tau - \theta_{j,t} - \theta_{\hat{j}_t,t}\}, \qquad \text{where } \hat{j}_t = \text{argmax}_j |\hat{\Delta}_{j,t}|, \\ \hat{H}_t &= \{i \,:\, |\hat{\Delta}_{i,t}| \,\geq\, |\hat{\Delta}_{\hat{j}_t,t}| - \tau + \theta_{j,t} + \max_{j \in \hat{C}_t} \theta_{j,t}\}, \\ \hat{B}_t &= \left\{i : |\hat{\Delta}_{\hat{j}_t,t}| - \tau - \theta_{j,t} - \theta_{\hat{j}_t,t} \,\leq\, |\hat{\Delta}_{i,t}| \,<\, |\hat{\Delta}_{\hat{j}_t,t}| - \tau + \theta_{j,t} + \max_{j \in \hat{C}_t} \theta_{j,t}\right\}. \end{aligned}$$

The pseudocode is given in Algorithm 3. Notice that the query condition defining Z_t now depends on an *average threshold* $\theta_{S \cup \hat{H}_t, t} = \frac{1}{|S \cup \hat{H}_t|} \sum_{j \in S \cup \hat{H}_t} \theta_{j,t}$.

3.4 Analysis, Second Version

The following theorem bounds the regret and the total number of queries issued by the second version of our algorithm, with high probability. The proof is similar to the proof of Theorem 12. We keep the definitions of the sets $H_{\varepsilon,t}$ and $B_{\varepsilon,t}$ as given in Section 3.2, but in the bound on N_T in

Theorem 17, we replace T'_ε with the more refined quantity T''_ε, defined as

$$T''_\varepsilon \;=\; \sum_{t=1}^{T} \frac{|H_{\varepsilon,t} \cup B_{\varepsilon,t}|}{K}\, \mathbf{1}\big\{|\Delta_t| > \varepsilon\big\}\, \mathbf{1}\big\{\exists S \subseteq B_{\varepsilon,t} \;:\; \Delta_t \Delta_{S \cup H_{\varepsilon,t},t} < 0 \;\vee\; |\Delta_{S \cup H_{\varepsilon,t},t}| \le \varepsilon\big\}\;.$$

Note that T''_ε is similar to T'_ε except that while T'_ε only counts the number of times that perturbations to the $\Delta_{j,t}$'s lead to conflict or low confidence predictions, T''_ε also accounts for the fraction of confident teachers involved in the conflict. We state the following bound on regret and on the overall number of queries.

Theorem 17 *Assume Algorithm 3 is run with a confidence parameter $\delta > 0$. Then with probability at least $1-\delta$ it holds for all $T \ge 3$ that*

$$\begin{aligned}
R_T &\le \inf_{\varepsilon>0}\left\{\varepsilon T_\varepsilon + T'_\varepsilon + O\left(\frac{K \log|A_T| \log(KT/\delta) + K\log^2|A_T|}{\varepsilon^2}\right)\right\} \\
&= \inf_{\varepsilon>0}\left\{\varepsilon T_\varepsilon + T'_\varepsilon + O\left(\frac{K d^2 \log^2(KT/\delta)}{\varepsilon^2}\right)\right\}\;, \\
N_T &\le K \inf_{\varepsilon>0}\left\{T_\varepsilon + T''_\varepsilon + O\left(\frac{K \log|A_T| \log(KT/\delta) + K\log^2|A_T|}{\varepsilon^2}\right)\right\} \\
&= K \inf_{\varepsilon>0}\left\{T_\varepsilon + T''_\varepsilon + O\left(\frac{K d^2 \log^2(KT/\delta)}{\varepsilon^2}\right)\right\}\;.
\end{aligned}$$

The bounds above resemble the bounds stated in Theorem 12 for the first version of the algorithm; all of these bounds contain two kinds of terms: hardness terms (T_ε, T'_ε, and T''_ε) and large deviation terms ($d \log T$-like factors). The regret bound for the second version of the algorithm is strictly inferior to the regret bound for the first version, as an additional factor of K multiplies the large deviation term. However, the bounds on the number of queries of the two algorithms are not directly comparable. On one hand, if a typical example only has a few confident teachers, we expect T''_ε to be much smaller than T'_ε, which could make the bound on N_T in Theorem 17 much smaller than its counterpart in Theorem 12. On the other hand, the bound in Theorem 17 has an additional factor of K multiplying its large deviation term.

As in the proofs of Theorem 2 and Theorem 12, to analyze the regret and number of queries made by the algorithm, we start by decomposing these terms. To decompose the regret, we note that Lemma 13 applies as before, and we have that for any $\varepsilon > 0$,

$$R_T \;\le\; \varepsilon T_\varepsilon + T'_\varepsilon + U_T + Q_{T,\varepsilon},$$

where T_ε is as defined in Equation (9), T'_ε is as defined in Equation (10), and U_T and $Q_{T,\varepsilon}$ are defined in Equation (11). To decompose the total number of queries, we require a new lemma.

Lemma 18 *If $(\Delta_{j,t} - \hat{\Delta}_{j,t})^2 \le \theta^2_{j,t}$ holds for all $j \in [K]$ and $t \in [T]$, then for any $\varepsilon > 0$, it holds that*

$$\begin{aligned}
N_T &\le K\left(T_\varepsilon + T''_\varepsilon + O\left(\frac{\sum_{j=1}^K \left(\log|A_{j,T}| \log(KT/\delta) + \log^2|A_{j,T}|\right)}{\varepsilon^2}\right)\right) \\
&\le K\left(T_\varepsilon + T''_\varepsilon + O\left(\frac{Kd^2 \log^2(KT/\delta)}{\varepsilon^2}\right)\right)\;.
\end{aligned}$$

Once again, proofs are given in Appendix C. We are left with the task of bounding U_T and $Q_{T,\varepsilon}$.

Lemma 19 *If* $(\Delta_{j,t} - \hat{\Delta}_{j,t})^2 \leq \theta_{j,t}^2$ *holds for all* $j \in [K]$ *and* $t \in [T]$*, then*

$$Q_{T,\varepsilon} = O\left(\frac{\sum_{j=1}^{K}\left(\log|A_{j,T}|\,\log(KT/\delta) + \log^2|A_{j,T}|\right)}{\varepsilon^2}\right) = O\left(\frac{K\,d^2\,\log^2(KT/\delta)}{\varepsilon^2}\right) \quad .$$

Lemma 20 *If* $(\Delta_{j,t} - \hat{\Delta}_{j,t})^2 \leq \theta_{j,t}^2$ *for all* $j \in [K]$ *and* $t \in [T]$*, then* $U_T = 0$*.*

Proofs of these lemmas are also given in Appendix C. As before, these lemmas hold under the condition that $(\Delta_{j,t} - \hat{\Delta}_{j,t})^2 \leq \theta_{j,t}^2$ for all $t \in [T]$ and $j \in [K]$. Again as done previously, a straightforward union bound over Lemma 7 in Section 2 applied to each of the K teachers verifies that this condition holds with high probability which in turn concludes the proof of Theorem 17.

Remark 21 *It should be clear that a low noise analysis, akin to the one presented in Sections 2.4 and 2.5 can be attempted, once low noise conditions in the vein of Tsybakov (2004) are formulated which take into account both the conflicting region defining* T_ε*, and the unstable regions defining* T'_ε*, and* T''_ε*. Rather than presenting explicit theoretical results of this sort here, we do prefer quantifying the label saving capability implied by teacher aggregation by the nontrivial experimental results contained in the next section.*

4. Experiments in the Multiple Teacher Setting

We report on the results of an empirical study carried out on a medium-size real-world data set. The goal of our experiments is to validate the theory and to quantify the effectiveness of our multiple teacher query selection mechanism in different multiple-teacher scenarios. Due to our difficulty in finding genuine multiple-teacher data sets of a significant size, we resorted to *simulating* the teachers through learning. This also allowed us to obtain a much more controlled experimental setup.

4.1 Data Set and Tasks

Our data are taken from a subset of the learning-to-rank data set MSLR-WEB10K.[7] This data set is a collection of (anonymized) query-url pairs collected from a commercial search engine (Microsoft Bing). Each query-url pair is represented by a feature vector and a human-generated relevance label between 0 (irrelevant) to 4 (perfectly relevant). Each feature vector is made up of 136 real or integer valued features.[8] MSLR-WEB10K is partitioned into five subsets, named S1 through S5: we only used S1 in our experiments. The S1 subset contains 241988 query-url pairs, with 2000 distinct queries, and about 121 urls per query (with a maximum of 809 urls and a minimum of 1 url per query). As a preprocessing step, we randomly shuffled the examples within each query and normalized the feature vectors to unit length.

We generated a binary classification data set by assigning the binary label "-1" to all query-url pairs with a relevance label of 0 and the binary label "+1" to all remaining pairs. This gave rise to a data set with balanced classes (roughly, 48% positive and 52% negative). We then simulated

7. Available at `http://research.microsoft.com/en-us/projects/mslr`.
8. After a quick scrutiny of the semantics of the features, we decided to drop features 126 through 131. Hence, we ended up with 130 usable features.

four different multiple-teacher scenarios, distinguished by the number of teachers involved ("few" or "many") and the amount of overlap between expertise regions ("nonoverlapping teachers" vs. "overlapping teachers").

This binary classification data set simply provides a binary label per example, and does not specify the identity of the teacher that provided that label. We simulated multiple teachers as follows: we grouped queries together in various ways (see below) and trained a linear classifier using half of the urls associated with each query in the group. The training was done using a single random-order pass of a full-information second-order Perceptron algorithm (Cesa-Bianchi et al., 2005a). The result is a linear classifier per query-group: we view each of these linear classifiers as a teacher. The specific subset of training queries in each group determines the expertise region of the respective teacher. The 119507 query-url pairs that were not used to simulate the teachers were later used to test our algorithm.

We defined the query groups in four different ways, to simulate four different multiple-teacher scenarios.

- **Few nonoverlapping teachers.** We generated 5 teachers by partitioning the 2000 queries into 5 sets (the first teacher is defined by (half of) the first 400 queries, the second teacher by (half of) the second 400 queries, and so on). Hence each teacher has acquired expertise in the subset of 400 queries seen during training.

- **Few overlapping teachers.** We generated 5 teachers by defining 5 overlapping sets of queries. Specifically, the first 500 queries are common to all teachers, and the remaining 1500 queries are partitioned equally among the teachers. Hence, each teacher is trained on examples from 800 queries.

- **Many nonoverlapping teachers.** We generated 100 teachers by partitioning the queries into 100 disjoint sets, each containing 20 queries. The resulting teachers turned out to be quite unreliable; some of them gave labels that were not far from random guessing at test time.

- **Many overlapping teachers.** We generated 100 teachers with partially overlapping expertise. All teachers share the first 100 queries, and the remaining 1900 queries are partitioned equally. Hence, each teacher is trained on examples from 100+19 = 119 queries.

Due to the variance introduced by the randomized training/test splits and the random order in which training examples were presented to the second-order Perceptron, we repeated the above process 10 times per scenario and averaged the results.

The reader should observe that the way we generated teachers makes our results comparable even across scenarios. In fact, despite the actual training/test split differs over scenarios, in all four scenarios and for all 2000 queries, half the urls (and associated labels) are used for training and half are used for test. So, in a sense, the set of teachers we generated in all scenarios collectively encode the same amount of information. That is, the data used for training the teachers are of the same size and query mixture across scenarios.

4.2 Algorithm and Baselines

On any given scenario, a teacher is then just a linear-threshold function. We generated teachers' opinions on the test set just by evaluating such functions on the test set instances. Table 1 gives

SCENARIO	BEST	WORST	AVG	STDDEV
FEW NONOVERLAPPING TEACHERS	19.9%	31.7%	24.9%	4.6%
FEW OVERLAPPING TEACHERS	20.5%	29.6%	24.5%	3.6%
MANY NONOVERLAPPING TEACHERS	16.3%	54.8%	25.5%	7.4%
MANY OVERLAPPING TEACHERS	17.0%	42.3%	24.5%	5.1%

Table 1: Performance (test set mistake rate) of the generated teachers in the four simulated scenarios. Results are averaged over 10 repetitions. "best", "worst", and "avg" are the (average) mistake rate of the best, worst and average performing teacher, respectively. "stddev" is the standard deviation of the mistake rates, and gives an idea of the difference in performance *across teachers* (not across repetitions).

relevant statistics about the teachers' performance on the test set (notice that such figures can only be computed after knowing the true labels on the test set—this information was never made available to the multiple teacher algorithm). As expected, best and worst teachers are farther apart in the nonoverlapping scenarios (correspondingly, "stddev" figures are larger), with a larger variability in the many teacher settings. Moreover, throughout the 10 repetitions, it often happened that among the many poorly trained classifiers (as are those produced within the "many nonoverlapping" setting), a few of them turned out to be significantly accurate on the test set. Likewise, some of them happened to be even worse than random guessing.

After simulating the teachers, we implemented a simplified version of our second-version multiple teacher algorithm (Algorithm 3), where the thresholds $\theta^2_{j,t}$ are simplified to

$$\theta^2_{j,t} = \alpha \mathbf{x}_t^\top A^{-1}_{j,t-1} \mathbf{x}_t \log(1+t), \tag{12}$$

and $\alpha > 0$ is a tunable parameter (independent of j and t). Hence our algorithm now has two parameters: $\tau \in [0,1]$ and $\alpha > 0$. The reason for this simplified $\theta_{j,t}$ is that the actual expression for $\theta_{j,t}$, as it appears in Algorithm 3, is the one suggested by the theory after significant mathematical overapproximations (large deviations, Hölder's inequality, etc.). This suggests that the exact expression for $\theta_{j,t}$ given in the pseudocode may be too conservative to work well in practice. In any event, observe that the factor $\alpha \log(1+t)$ in (12) is a good proxy for the factor $1+4\sum_{i=1}^{t-1} Z_i r_i + 36\log(Kt/\delta)$ in the algorithm's pseudocode, once we let α range over the positive reals.

The following three baselines were used in our comparative study.

- BEST TEACHER in hindsight on the test set. This is the predictor that would be learned on-the-fly by a standard expert algorithm (e.g., Weighted Majority—see Littlestone and Warmuth, 1994; Cesa-Bianchi and Lugosi, 2006), where teachers are experts, and the algorithm has at its disposal both the true labels of the test set and the prediction of all teachers. Recall that the true labels of the test set are not available to our algorithm. Because this algorithm is expected to make at least as many mistakes as the best expert, the "best" column in Table 1 delivers optimistic approximations to the actual performance of this algorithm in the four scenarios. The associated number of queries made to the teachers is the largest possible, that is, the size of the test set (119507) times the number of teachers (119507×5 in the "few teacher" scenarios, and 119507×100 in the "many teacher" scenarios).

SCENARIO	BEST TEACH.	FLAT MAJORITY	FULL-INFO ALG. 3
FEW NONOVERLAPPING TEACHERS	19.9%	16.7%	15.8%
FEW OVERLAPPING TEACHERS	20.5%	17.9%	15.8%
MANY NONOVERLAPPING TEACHERS	16.3%	15.6%	15.6%
MANY OVERLAPPING TEACHERS	17.0%	15.7%	15.7%

Table 2: Performance (test set mistake rate) of the three tested baselines in the four simulated scenarios. Results are averaged over 10 repetitions. The "Best Teacher" figures are taken from Table 1.

- FLAT MAJORITY of teachers. This algorithm asks all teachers and predicts with their flat majority.[9] Like the BEST TEACHER baseline, this algorithm queries all of the teachers all the time. Unlike BEST TEACHER, this algorithm does not receive any feedback on the true labels of the test set.

- FULL-INFORMATION version of our second-version Algorithm (Algorithm (3)). This is our algorithm with $\theta^2_{j,t}$ fixed to the value ∞ for all j and t. Since $\theta^2_{j,t} = \infty$ implies $Z_t = 1$ and $\hat{C}_t = [K]$ for all t (thereby making τ immaterial), this algorithm predicts by aggregating all teachers via a margin-based majority and, as before, querying all labels from all teachers. Again, no ground-truth feedback is given. Hence, this baseline is just a weighted version of FLAT MAJORITY, where the weights are given by the estimated margins $\hat{\Delta}_{j,t}$ computed by Algorithm 3 operating in a "full information" mode.

4.3 Results and Comments

We measured the error rate on the test set and the average number of requested labels per example. Figure 1 shows test error rate as a function of the per-teacher query rate (i.e., the average fraction of times we query the teachers). The figure displays the test error rate of our algorithm compared to the three baselines mentioned above, in each of the four scenarios, with $\tau = 0.3$ and different values of α in $[0.01, 10]$. Very similar plots are obtained for other values of τ.[10] Increasing α causes a steady increase in the (average) per-teacher query rate, but surprisingly enough, has a negligible effect on test error rate across most of its range (hence the flattish plots in Figure 1). In particular, a query rate of about 1% is already sufficient to get very close to the smallest test error rate achieved by the algorithm. As for comparison to the baselines, the following comments can be made.

- Our algorithm significantly outperforms all baselines in the "few teacher" scenarios, but is about the same as the two majority baselines in the "many teacher" scenarios. Notice, however, that this comparison is unfairly penalizing our algorithm in that the baselines do achieve their results by asking all of the teachers all of the time. Moreover, it is worth stressing that

9. Alternatively, this algorithm picks a teacher uniformly at random and goes with its label. In our experiments, we did not test this randomized version due to the high variance of the results, especially in the "many teacher" scenarios—see the last two rows in Table 1.

10. For instance, in the "few nonoverlapping" setting, when $\tau = 0.0$ and α ranges over [0.01,10] the test error rate of our algorithm ranges between 15.5% and 15.7%; when $\tau = 0.7$ the test error ranges between 15.6% and 15.7%. In the "many overlapping" setting, when $\tau = 0.0$ and $\alpha \in [0.01,10]$ we obtain a test error between 15.6% and 15.9%; When $\tau = 0.7$, the range is between 15.4% and 15.5%.

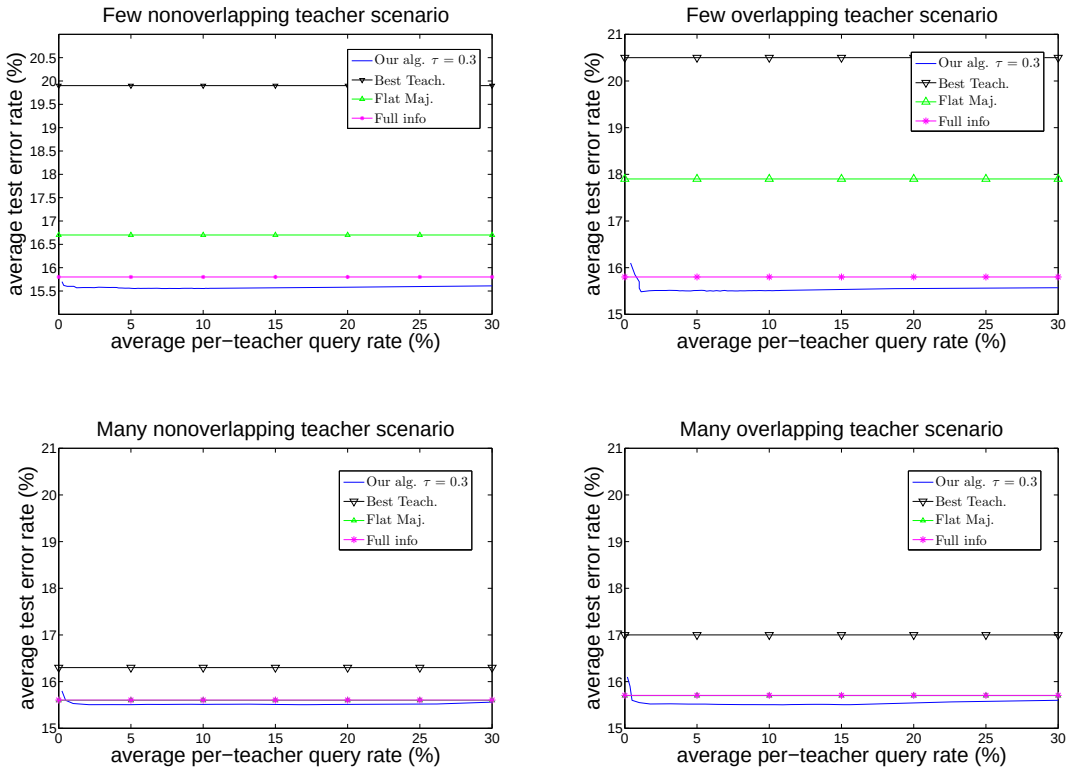

Figure 1: Average per-teacher query rate vs. test error rates in the four scenarios. Results are averaged over 10 repetitions. The query rate of the multiple selective sampler - second version ("Our alg.") is obtained by setting $\tau = 0.3$, and letting α in (12) vary across the range $[0.01, 10]$. In any given scenario, the per-teacher query rate of our algorithm is the average fraction of labels requested to the teachers out of the total number of labels available in that scenario. For instance, an average per-teacher query rate of 10% achieved in a few teacher scenario means that, averaged over the 10 repetitions, the total no. of queries made to the five teachers was $119507 \times 5 \times 10\% = 59753.5$. Hence, each of the 5 teachers received on average 11950.7 queries. The test error rates of the three baselines are taken from Table 2, and are plotted (as horizontal lines) just for reference.

though our plots display average rates over repetitions (i.e., over train/test splits), the above comparative behavior did consistently occur in *every* single repetition.

- We found somewhat surprising that FULL-INFO does not improve on FLAT MAJORITY. Moreover, since FULL-INFO can be obtained by our algorithm, just by setting $\alpha = \infty$ in (12), we see that more teacher labels can even be detrimental. This phenomenon is statistically significant only in the "few teacher" scenarios.

- The more teachers we have at our disposal, the more beneficial is the process of averaging over them. Notably, in our data set, many unreliable (but nonoverlapped) teachers queried all the times and aggregated by a flat average (aka FLAT MAJORITY) is about as good in terms

of accuracy as running more sophisticated weighted averages. Still, our experiments show that there is no need to query all of the teachers in order to achieve this accuracy.

- Aggregating opinions of teachers with a good amount of overlapping expertise (as in the two "overlapping teacher" settings), might be detrimental, as evinced by comparing the first row in Table 2 to the second one, and the third to the fourth one. Similar conclusions are suggested by the behavior of our algorithm as presented in Figure 1.

Finally, we make a few comments on the role of the parameter τ in our algorithm. As mentioned above, we observed that the value of τ does not have a significant influence on the algorithm's test error rate or label query rate. In a sense, this is a lucky circumstance, since we initially expected the tuning of τ to be a nontrivial task.[11] The value of τ does however play an important role in the "degree of aggregation" of teachers: When $Z_t = 1$, setting τ close to 0 makes the algorithm query only the (estimated) most confident teacher at time t, whereas setting τ close to 1 causes the algorithm to query all teachers. For instance, in the "many nonoverlapping teachers" scenario, if $\tau = 0.3$ (as in the plot in Figure 1 (d)), and $\alpha = 0.1$, the most queried teacher receives 5440 queries (out of 119507), and the least queried teacher receives 3319. In the same scenario with $\tau = 0.0$ and the same value of α, the most queried teacher receives 10849 queries while the least queried teacher gets only 1050. Hence, our algorithm exhibits a desirable fine-grained selection capability of the subsets of teachers to query, thereby making it significantly different from the all-or-none strategy followed by the first version of our algorithm (Algorithm 2), which we do not expect to work as well in practice.

5. Conclusions and Open Questions

We introduced a new algorithm in the online selective sampling framework, where instances are chosen by an adaptive adversary and labels are sampled from a linear stochastic model. We gave sharp bounds on the regret and on the number of queries made by this algorithm, improving over previous algorithms and closing some important open questions on this topic. The same machinery can also be used to build efficient active learning algorithms working under standard statistical assumptions. We then lifted the above to the more involved setting where multiple unreliable teachers are available. We presented two algorithms and corresponding analyses. We concluded with a preliminary empirical study that demonstrates how the second version of our algorithm outperforms various intuitive baselines, both in terms of accuracy and total number of queries.

We leave some open problems for future research: The bound on N_T in Theorem 2 is tight w.r.t. ε (see the lower bound by Cesa-Bianchi et al., 2009), but need not be tight w.r.t. d. This might be due to the way we constructed our martingale argument to prove Lemma 7. Resolving this issue remains an open problem. Second, it would be interesting to generalize our results to other stochastic label models, such as logistic models, and to understand how closely each model matches the true behavior of human teachers. Third, the bounds in the multiple teacher setting (Theorems 12 and 17) are likely to be suboptimal, and might perhaps be improved by exploiting the interaction structure among teachers. Fourth, it would be interesting to extend our work to a setting where different teachers charge different rates. For example, one could imagine a setting where the cost of each label depends on each teacher's confidence in his own answer. This setting is closer to the

11. Consider that the absence of ground-truth feedback makes standard cross-validation techniques somewhat problematic.

proactive learning setting (Donmez and Carbonell, 2008; Yang and Carbonell, 2009a,b). These and other open problems provide many opportunities for interesting future research on this topic.

Acknowledgments

We thank the Action Editor for his timely handling of this paper. We also thank the anonymous reviewers for their helpful comments. This research was done while the second and the third authors were visiting Microsoft Research at Redmond. The second author acknowledges the PASCAL2 Network of Excellence under EC grant 216886 for supporting travel expenses to the conference.

Appendix A.

This appendix contains the large deviation inequalities we use throughout the paper.

Lemma 22 *(Kakade and Tewari, 2008)*
Suppose $X_1, X_2, ..., X_T$ is a martingale difference sequence with $|X_t| \leq b$. Let $\mathrm{Var}_t(X_t) = \mathrm{Var}(X_t | X_1, ..., X_{t-1})$, and $V = \sum_{t=1}^{T} \mathrm{Var}_t(X_t)$. Then for any $\delta < 1/e$ and $T \geq 3$, we have

$$P\left(\sum_{t=1}^{T} X_t > \max\left\{ \sqrt{4V \log \frac{4\log T}{\delta}}, 3b \log \frac{4\log T}{\delta} \right\} \right) \leq \delta.$$

Lemma 23 *With the notation introduced in Section 2, define*

$$\mu_t = \sum_{i=1}^{t} Z_i (\Delta_i - \hat{\Delta}'_i)^2, \qquad \Sigma_t = \sum_{i=1}^{t} Z_i \left((y_i - \hat{\Delta}'_i)^2 - (y_i - \Delta_i)^2 \right).$$

Assume that Selective Sampler in Section 2 is run with confidence parameter $\delta \in (0,1]$, and let $t \geq 3$. Then

(i) with probability at least $1 - \delta/t^2$ we have $\mu_t \leq 2\Sigma_t + 144 \log(t/\delta)$;

(ii) with probability at least $1 - \delta/t^2$ we have $-\frac{1}{2}\Sigma_t \leq 36 \log(t/\delta)$.

Proof Set $M_i = Z_i (\Delta_i - y_i)(\Delta_i - \hat{\Delta}'_i)$, and observe that M_i can be rewritten as

$$M_i = \frac{1}{2} Z_i \left((\Delta_i - \hat{\Delta}'_i)^2 - \left((y_i - \hat{\Delta}'_i)^2 - (y_i - \Delta_i)^2 \right) \right),$$

which implies $\frac{1}{2}(\mu_t - \Sigma_t) = \sum_{t=1}^{t} M_i$. Now, $M_1, ..., M_t$ is a martingale difference sequence w.r.t. history and current $\mathbf{x}_i$. This is because $\mathbf{E}_i[M_i] = Z_i (\Delta_i - \mathbf{E}_i[y_i])(\Delta_i - \hat{\Delta}'_i) = 0$. Since $|\Delta_t|, |\hat{\Delta}_t| \leq 1$, we also have that $|M_i| \leq 4$. Let $\mathrm{Var}_i(\cdot)$ denote the conditional variance $\mathrm{Var}(\cdot \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_i, y_1, \ldots, y_{i-1})$. Observing that

$$\mathrm{Var}_i(M_i) = Z_i (\Delta_i - \hat{\Delta}'_i)^2 \mathrm{Var}_i\left((\Delta_i - y_i)^2 \right) \leq \frac{4}{3} Z_i (\Delta_i - \hat{\Delta}'_i)^2$$

holds, an application of Lemma 22 yields

$$\frac{1}{2}(\mu_t - \Sigma_t) \leq \max\left\{ \sqrt{6\,\mu_t \,\log\left(\frac{4t^2 \log t}{\delta} \right)}, 12 \log\left(\frac{4t^2 \log t}{\delta} \right) \right\}. \tag{13}$$

We now use the inequality $\sqrt{ab} \leq \frac{a+b}{2}$ to (13) with $a = \mu_t/2$ and $b = 12\log\left(\frac{4t^2\log t}{\delta}\right)$. This implies

$$\frac{1}{2}\left(\mu_t - \Sigma_t\right) \leq \mu_t/4 + 12\log\left(\frac{4t^2\log t}{\delta}\right)$$

which in turn implies (i). To prove (ii), we again apply $\sqrt{ab} \leq \frac{a+b}{2}$ to (13), this time with $a = \mu_t$ and $b = 6\log\left(\frac{4t^2\log t}{\delta}\right)$. ■

Appendix B.

Most of the steps in the proofs of these lemmas appear in the papers by Azoury and Warmuth (2001) and Cesa-Bianchi et al. (2005a). The proofs are provided here for completeness.

Lemma 24 *With the notation introduced in Section 2, we have that for each $t = 1, 2, \ldots$ the following inequalities hold :*

(i) $\mathbf{x}^\top A_{t-1}\mathbf{x}_t \leq 2r_t$;

(ii) $Z_t r_t \leq \log\frac{|A_t|}{|A_{t-1}|}$;

(iii) $\sum_{i=1}^t Z_i r_i \leq \log|A_t| \leq d\log(1+N_t) = O(d\log t)$.

Proof To prove (i), note that on the rounds we do not query, $A_t = A_{t-1}$ and so $r_t = \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t$. On the rounds we do query, $A_t = A_{t-1} + \mathbf{x}_t\mathbf{x}_t^\top$, and so by the matrix inversion formula

$$A_t^{-1} = A_{t-1}^{-1} - \frac{A_{t-1}^{-1}\mathbf{x}_t\mathbf{x}_t^\top A_{t-1}^{-1}}{1+\mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t}$$

we see that

$$r_t = \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t - \frac{(\mathbf{x}_t^\top A_{t-1}\mathbf{x}_t)^2}{1+\mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t}\ .$$

This automatically gives us that $r_t \leq \mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t$. Further since $\mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t \leq 1$, we can conclude that $r_t \geq \frac{1}{2}\mathbf{x}_t^\top A_{t-1}^{-1}\mathbf{x}_t$. Hence we conclude that for any t, $r_t \leq \mathbf{x}^\top A_{t-1}\mathbf{x}_t \leq 2r_t$.

Now to prove (ii), note that since whenever we query, $A_t = A_{t-1} + \mathbf{x}_t\mathbf{x}_t^\top$, using the identity, $\mathbf{x}_t^\top(A_{t-1} + \mathbf{x}_t\mathbf{x}_t^\top)^{-1}\mathbf{x}_t = 1 - \frac{|A_t|}{|A_{t-1}|}$ and the fact that $1 - x \leq \log(x)$, we see that

$$r_t \leq \log\frac{|A_t|}{|A_{t-1}|}.$$

To get (iii), we sum up and resolve the telescoping sum as

$$\sum_{i=1}^t Z_i r_i \leq \sum_{i=1}^t Z_i \log\frac{|A_i|}{|A_{i-1}|} = \log|A_t| \leq d\ \log(1+N_t) = O(d\ \log t)\ .$$

■

Lemma 25 *With the notation introduced in Section 2, the following holds for any* $\mathbf{u} : ||\mathbf{u}|| \leq 1$*:*

(i) If t *is such that* $Z_t = 1$ *we have*

$$\frac{1}{2}\left((y_t - \mathbf{w}'_{t-1}{}^{\top}\mathbf{x}_t)^2 - (y_t - \mathbf{u}^{\top}\mathbf{x}_t)^2\right) \;=\; d_{t-1}(\mathbf{u}, \mathbf{w}'_{t-1}) - d_t(\mathbf{u}, \mathbf{w}_t) + d_t(\mathbf{w}'_{t-1}, \mathbf{w}_t)\;;$$

(ii) If t *is such that* $Z_t = 1$ *we have* $d_t(\mathbf{w}'_{t-1}, \mathbf{w}_t) \leq 2\,r_t$;

(iii) If t *is such that* $Z_t = 1$ *we have* $d_t(\mathbf{u}, \mathbf{w}'_t) \leq d_t(\mathbf{u}, \mathbf{w}_t)$;

(iv) For any $t = 1, 2, \ldots$*, we have*

$$\frac{Z_t}{2}\left((y_t - \mathbf{w}'_{t-1}{}^{\top}\mathbf{x}_t)^2 - (y_t - \mathbf{u}^{\top}\mathbf{x}_t)^2\right) \leq Z_t\,\left(d_{t-1}(\mathbf{u}, \mathbf{w}'_{t-1}) - d_t(\mathbf{u}, \mathbf{w}'_t)\right) + 2\log\frac{|A_t|}{|A_{t-1}|}.$$

Proof To prove (i), define $\alpha_t := d_{t-1}(\mathbf{u}, \mathbf{w}'_{t-1}) - d_t(\mathbf{u}, \mathbf{w}_t) + d_t(\mathbf{w}'_{t-1}, \mathbf{w}_t)$. Using the definition of d_t, we have that

$$\alpha_t \;=\; \frac{1}{2}\mathbf{u}^{\top}(A_{t-1} - A_t)\mathbf{u} + \mathbf{u}^{\top}(A_t\mathbf{w}_t - A_{t-1}\mathbf{w}'_{t-1}) + \frac{1}{2}\mathbf{w}'_{t-1}(A_{t-1} + A_t)\mathbf{w}'_{t-1} - \mathbf{w}'_{t-1}A_t\mathbf{w}_t \;\;.$$

Using the recursive definition $A_t = A_{t-1} + \mathbf{x}_t\mathbf{x}_t^{\top}$ and rearranging terms in the right-hand side above, we get

$$\begin{aligned}\alpha_t \;&=\; \frac{1}{2}\left((\mathbf{w}'_{t-1})^{\top}\mathbf{x}_t\mathbf{x}_t^{\top}\mathbf{w}'_{t-1} - \mathbf{u}^{\top}\mathbf{x}_t\mathbf{x}_t^{\top}\mathbf{u}\right) + (\mathbf{u}^{\top} - \mathbf{w}'_{t-1})(A_t\mathbf{w}_t - A_{t-1}\mathbf{w}'_{t-1}) \\ &=\; \frac{1}{2}\left((\mathbf{w}'_{t-1}{}^{\top}\mathbf{x}_t)^2 - (\mathbf{u}^{\top}\mathbf{x}_t)^2\right) + (\mathbf{u}^{\top} - \mathbf{w}'_{t-1})(A_t\mathbf{w}_t - A_{t-1}\mathbf{w}'_{t-1}) \;\;.\end{aligned}$$

By definition, $A_t\mathbf{w}_t = A_{t-1}\mathbf{w}'_{t-1} + y_t\mathbf{x}_t$. Plugging this equality into the equation above gives

$$\begin{aligned}\alpha_t \;&=\; \frac{1}{2}\left((\mathbf{w}'_{t-1}{}^{\top}\mathbf{x}_t)^2 - (\mathbf{u}^{\top}\mathbf{x}_t)^2\right) + y_t(\mathbf{u}^{\top} - \mathbf{w}'_{t-1})\mathbf{x}_t \\ &=\; \frac{1}{2}\left((y_t - \mathbf{w}'_{t-1}{}^{\top}\mathbf{x}_t)^2 - (y_t - \mathbf{u}^{\top}\mathbf{x}_t)^2\right)\;,\end{aligned}$$

thereby proving (i).

To prove (ii), we rewrite $d_t(\mathbf{w}'_{t-1}, \mathbf{w}_t)$ as

$$\begin{aligned}d_t(\mathbf{w}'_{t-1}, \mathbf{w}_t) \;&=\; \frac{1}{2}(\mathbf{w}'_{t-1} - \mathbf{w}_t)^{\top}A_t(\mathbf{w}'_{t-1} - \mathbf{w}_t) \\ &=\; \frac{1}{2}(A_t\mathbf{w}'_{t-1} - A_t\mathbf{w}_t)^{\top}A_t^{-1}(A_t\mathbf{w}'_{t-1} - A_t\mathbf{w}_t) \;\;.\end{aligned}$$

Using $A_t\mathbf{w}_t = A_{t-1}\mathbf{w}'_{t-1} + y_t\mathbf{x}_t$, the above becomes

$$d_t(\mathbf{w}'_{t-1}, \mathbf{w}_t) \;=\; \frac{1}{2}\left((A_t - A_{t-1})\mathbf{w}'_{t-1} - y_t\mathbf{x}_t\right)^{\top}A_t^{-1}\left((A_t - A_{t-1})\mathbf{w}'_{t-1} - y_t\mathbf{x}_t\right) \;\;.$$

Using $A_t = A_{t-1} + \mathbf{x}_t \mathbf{x}_t^\top$, we have

$$\begin{aligned} d_t(\mathbf{w}'_{t-1}, \mathbf{w}_t) &= \frac{1}{2}\left(\mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}'_{t-1} - y_t \mathbf{x}_t\right) A_t^{-1} \left(\mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}'_{t-1} - y_t \mathbf{x}_t\right) \\ &= \frac{(\mathbf{w}'_{t-1}{}^\top \mathbf{x}_t - y_t)^2}{2} \mathbf{x}_t^\top A_t^{-1} \mathbf{x}_t \\ &= \frac{(\mathbf{w}'_{t-1}{}^\top \mathbf{x}_t - y_t)^2}{2} r_t \\ &= \frac{(\hat{\Delta}'_t - y_t)^2}{2} r_t \\ &\le 2 r_t \ , \end{aligned}$$

where the last step uses $|\hat{\Delta}'_t| \le 1$

To prove (iii) we observe that, $\mathbf{w}'_t$, as defined in Algorithm 1, is the projection of $\mathbf{w}_t$ onto the convex set $C_t = \{\mathbf{w} : |\mathbf{w}^\top \mathbf{x}_t| \le 1\}$ w.r.t. Bregman divergence d_t. By the theorem of generalized projections we have that

$$0 \le d_t(\mathbf{w}'_t, \mathbf{w}_t) \le d_t(\mathbf{u}, \mathbf{w}_t) - d_t(\mathbf{u}, \mathbf{w}'_t) \ .$$

holds for any $\mathbf{u} \in C_t$. Since C_t includes the unit ball $\{\mathbf{u} : ||\mathbf{u}|| \le 1\}$ the claim follows.

Finally, to prove (iv), observe that when t is such that $Z_t = 0$ then both sides of the inequality are 0 (since $A_t = A_{t-1}$). On the other hand, when $Z_t = 1$ we just combine (i), (ii), (iii), and Lemma 24 (ii) to give the required inequality. ■

Appendix C.

Proof sketch of Theorem 9. We rely on Theorem 2, where the role of T_ε is neatly handled by the low-noise assumption combined with a standard Chernoff bound. In particular, since $\mathbb{E}[T_\varepsilon] \le cT\varepsilon^\alpha$, we can easily conclude that for any $\delta > 0$, with probability at least $1-\delta$ over sample $\mathbf{x}_1, \ldots, \mathbf{x}_T$ we have $T_\varepsilon \le \frac{3c}{2} T\varepsilon^\alpha + O(\log(1/\delta))$. We optimize over ε the bounds on R_T and N_T contained in Theorem 2. We obtain that, with the same probability,

$$\begin{aligned} R_T &= O\left((d \log(T/\delta))^{\frac{\alpha+1}{\alpha+2}} \, T^{\frac{1}{\alpha+2}} + \log(1/\delta) \right) \ , \\ N_T &= O\left((d^2 \log^2(T/\delta))^{\frac{\alpha}{\alpha+2}} \, T^{\frac{2}{\alpha+2}} + \log(1/\delta) \right) \ . \end{aligned} \tag{14}$$

Now define

$$K_t = \left(P'_t(y_t \, \hat{\Delta}_t < 0) - P'_t(y_t \, \Delta_t < 0)\right) - \left(P_t(y_t \, \hat{\Delta}_t < 0) - P_t(y_t \, \Delta_t < 0)\right) \ ,$$

and note that $K_1, ..., K_T$ forms a martingale difference sequence. Let $\mathbf{E'}_t[\cdot]$ denote the conditional expectation $\mathbf{E}[\cdot \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, y_1, \ldots, y_{t-1}]$ and $\mathrm{Var'}_t(\cdot)$ be the conditional variance

$\mathrm{Var}(\cdot \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, y_1, \ldots, y_{t-1})$.[12] We have

$$\begin{aligned}
\mathrm{Var}'_t[K_t] &= \mathbf{E'}_t\left[K_t^2\right] \\
&\le 2\left(\left(P'_t(y_t\ \hat{\Delta}_t < 0) - P'_t(y_t\ \Delta_t < 0)\right)\right)^2 \\
&\quad + 2\,\mathbf{E'}_t\left[\left(P_t(y_t\ \hat{\Delta}_t < 0) - P_t(y_t\ \Delta_t < 0)\right)^2\right] \\
&\quad \text{(using } (a-b)^2 \le 2a^2 + 2b^2\text{)} \\
&\le 2\left(P'_t(y_t\ \hat{\Delta}_t < 0) - P'_t(y_t\ \Delta_t < 0)\right) + 2\,\mathbf{E'}_t\left[P_t(y_t\ \hat{\Delta}_t < 0) - P_t(y_t\ \Delta_t < 0)\right] \\
&\quad \text{(using } P'_t(y_t\ \hat{\Delta}_t < 0) \ge P'_t(y_t\ \Delta_t < 0) \text{ and } P_t(y_t\ \hat{\Delta}_t < 0) \ge P_t(y_t\ \Delta_t < 0)\text{)} \\
&= 4\left(P'_t(y_t\ \hat{\Delta}_t < 0) - P'_t(y_t\ \Delta_t < 0)\right)\ .
\end{aligned}$$

Following Lemma 22 and overapproximating we have that, with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{t=1}^{T}\left(P'_t(y_t\ \hat{\Delta}_t < 0) - P'_t(y_t\ \Delta_t < 0)\right) &\le 2\,R_T + O\left(\log\left(\frac{\log T}{\delta}\right)\right) \\
&= O\left((d\log(T/\delta))^{\frac{\alpha+1}{\alpha+2}}\ T^{\frac{1}{\alpha+2}} + \log\left(\frac{\log T}{\delta}\right)\right)\ ,
\end{aligned}$$

the last equality deriving from (14). Dividing by T concludes the proof.

Proof of Lemma 13. We upper bound each of the summands in Equation (8) individually. We begin as in Equation (3) in the proof of Lemma 3. This gives us

$$P_t(y\hat{\Delta}_t < 0) - P_t(y\Delta_t < 0)\ \le\ \varepsilon\,\mathbb{1}\left\{\Delta_t\hat{\Delta}_t < 0, |\Delta_t| \le \varepsilon\right\} + \mathbb{1}\left\{\Delta_t\hat{\Delta}_t < 0, |\Delta_t| > \varepsilon\right\}|\Delta_t|\ . \tag{15}$$

The first term on the right-hand side above is simply upper bounded by $\varepsilon\,\mathbb{1}\left\{|\Delta_t| \le \varepsilon\right\}$. To upper bound the second term, we recall that $|\Delta_t| \le 1$ and bound $\mathbb{1}\left\{\Delta_t\hat{\Delta}_t < 0, |\Delta_t| > \varepsilon\right\}$ by

$$\begin{aligned}
&\mathbb{1}\left\{\Delta_t\hat{\Delta}_t < 0, |\Delta_t| > \varepsilon\right\}\mathbb{1}\left\{\exists S \subseteq B_{\varepsilon,t}\ :\ \Delta_t\Delta_{S\cup H_{\varepsilon,t}} < 0\ \vee\ |\Delta_{S\cup H_{\varepsilon,t}}| \le \varepsilon\right\} \\
&\qquad + \mathbb{1}\left\{\Delta_t\hat{\Delta}_t < 0, |\Delta_t| > \varepsilon\right\}\mathbb{1}\left\{\forall S \subseteq B_{\varepsilon,t}\ :\ \Delta_t\Delta_{S\cup H_{\varepsilon,t}} \ge 0,\ |\Delta_{S\cup H_{\varepsilon,t}}| > \varepsilon\right\} \\
&\le\ \mathbb{1}\left\{|\Delta_t| > \varepsilon\right\}\mathbb{1}\left\{\exists S \subseteq B_{\varepsilon,t}\ :\ \Delta_t\Delta_{S\cup H_{\varepsilon,t}} < 0\ \vee\ |\Delta_{S\cup H_{\varepsilon,t}}| \le \varepsilon\right\} \\
&\qquad + \mathbb{1}\left\{\Delta_t\hat{\Delta}_t < 0\right\}\mathbb{1}\left\{\forall S \subseteq B_{\varepsilon,t}\ :\ \Delta_t\Delta_{S\cup H_{\varepsilon,t}} \ge 0,\ |\Delta_{S\cup H_{\varepsilon,t}}| > \varepsilon\right\} \\
&\le\ \mathbb{1}\left\{|\Delta_t| > \varepsilon\right\}\mathbb{1}\left\{\exists S \subseteq B_{\varepsilon,t}\ :\ \Delta_t\Delta_{S\cup H_{\varepsilon,t}} < 0\ \vee\ |\Delta_{S\cup H_{\varepsilon,t}}| \le \varepsilon\right\} \\
&\qquad + \bar{Z}_t\,\mathbb{1}\left\{\Delta_t\hat{\Delta}_t < 0\right\} + Z_t\,\mathbb{1}\left\{\forall S \subseteq B_{\varepsilon,t}\ :\ \Delta_t\Delta_{S\cup H_{\varepsilon,t}} \ge 0,\ |\Delta_{S\cup H_{\varepsilon,t}}| > \varepsilon\right\}\ .
\end{aligned} \tag{16}$$

We plug Equation (16) into the right-hand side of Equation (15) to obtain the desired upper-bound on $P_t(y\hat{\Delta}_t < 0) - P_t(y\Delta_t < 0)$. Summing over t completes the proof.

Proof of Lemma 14. It is straightforward to verify that

$$\begin{aligned}
Z_t\ &=\ Z_t\,\mathbb{1}\left\{|\Delta_t| \le \varepsilon\right\} + Z_t\,\mathbb{1}\left\{|\Delta_t| > \varepsilon\right\} \\
&\le\ Z_t\,\mathbb{1}\left\{|\Delta_t| \le \varepsilon\right\}\ +\ Z_t\,\mathbb{1}\left\{|\Delta_t| > \varepsilon\right\}\mathbb{1}\left\{\exists S \subseteq B_{\varepsilon,t}\ :\ \Delta_t\Delta_{S\cup H_{\varepsilon,t}} < 0\ \vee\ |\Delta_{S\cup H_{\varepsilon,t}}| \le \varepsilon\right\} \\
&\quad + Z_t\,\mathbb{1}\left\{|\Delta_t| > \varepsilon\right\}\mathbb{1}\left\{\forall S \subseteq B_{\varepsilon,t}\ :\ \Delta_t\Delta_{S\cup H_{\varepsilon,t}} \ge 0,\ |\Delta_{S\cup H_{\varepsilon,t}}| > \varepsilon\right\} \\
&\le\ \mathbb{1}\left\{|\Delta_t| \le \varepsilon\right\}\ +\ \mathbb{1}\left\{|\Delta_t| > \varepsilon\right\}\mathbb{1}\left\{\exists S \subseteq B_{\varepsilon,t}\ :\ \Delta_t\Delta_{S\cup H_{\varepsilon,t}} < 0\ \vee\ |\Delta_{S\cup H_{\varepsilon,t}}| \le \varepsilon\right\} \\
&\quad + Z_t\,\mathbb{1}\left\{\forall S \subseteq B_{\varepsilon,t}\ :\ \Delta_t\Delta_{S\cup H_{\varepsilon,t}} \ge 0,\ |\Delta_{S\cup H_{\varepsilon,t}}| > \varepsilon\right\}\ .
\end{aligned}$$

12. Notice the difference between the conditional expectation and conditional variance used here and those used in the proof of Lemma 23.

Summing over t proves the bound.

Proof of Lemma 15. First, note that, by the way Algorithm 2 is defined,

$$\begin{aligned}
Z_t &= \mathbf{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \hat{\Delta}_t \hat{\Delta}_{S\cup\hat{H}_t,t} < 0 \,\vee\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \le \theta_t\big\} \\
&= \mathbf{1}\big\{\exists S \subseteq \hat{B}_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \le \theta_t\big\} \\
&\quad + \mathbf{1}\big\{\forall S \subseteq \hat{B}_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| > \theta_t\big\}\, \mathbf{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \hat{\Delta}_t \hat{\Delta}_{S\cup\hat{H}_t,t} < 0\big\} \\
&\le \mathbf{1}\big\{\exists S \subseteq \hat{B}_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \le \theta_t\big\} \\
&\quad + \mathbf{1}\big\{|\hat{\Delta}_t| > \theta_t\big\}\, \mathbf{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \hat{\Delta}_t \hat{\Delta}_{S\cup\hat{H}_t,t} < 0,\ |\hat{\Delta}_{S\cup\hat{H}_t,t}| > \theta_t\big\}.
\end{aligned}$$

We focus on the second term on the right-hand side above. Using the assumption that $|\Delta_{j,t} - \hat{\Delta}_{j,t}| \le \theta_t$ for all $j \in [K]$ together with Jensen's inequality, we have that $|\hat{\Delta}_t - \Delta_{\hat{C}_t,t}| \le \theta_t$ and $|\hat{\Delta}_{S\cup\hat{H}_t,t} - \Delta_{S\cup\hat{H}_t,t}| \le \theta_t$ for any S. Now, if S is such that $\hat{\Delta}_t \hat{\Delta}_{S\cup\hat{H}_t,t} < 0$, $|\hat{\Delta}_{S\cup\hat{H}_t,t}| > \theta_t$, and $|\hat{\Delta}_t| > \theta_t$, then it also holds that $\Delta_{\hat{C}_t,t}\Delta_{S\cup\hat{H}_t,t} < 0$. Moreover, if there exists $S \subseteq \hat{B}_t$ such that $\Delta_{\hat{C}_t,t}\Delta_{S\cup\hat{H}_t,t} < 0$ then either $\Delta_t \Delta_{S\cup\hat{H}_t,t} < 0$ or $\Delta_t \Delta_{\hat{C}_t,t} < 0$. Since $\hat{C}_t = \hat{H}_t \cup \hat{B}_t$ we have that

$$\mathbf{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \Delta_{\hat{C}_t,t}\Delta_{S\cup\hat{H}_t,t} < 0\big\} \;\le\; \mathbf{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \Delta_t \Delta_{S\cup\hat{H}_t,t} < 0\big\}\,.$$

Putting together, we can write

$$Z_t \;\le\; \mathbf{1}\big\{\exists S \subseteq \hat{B}_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \le \theta_t\big\} + \mathbf{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \Delta_t \Delta_{S\cup\hat{H}_t,t} < 0\big\}\;.$$

Using the above, we can decompose Z_t as follows

$$\begin{aligned}
Z_t &= Z_t\, \mathbf{1}\big\{4\theta_t > \varepsilon\big\} + Z_t\, \mathbf{1}\big\{4\theta_t \le \varepsilon\big\} \\
&\le Z_t\, \mathbf{1}\big\{4\theta_t > \varepsilon\big\} + \mathbf{1}\big\{\exists S \subseteq \hat{B}_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \le \theta_t\big\}\, \mathbf{1}\big\{4\theta_t \le \varepsilon\big\} \\
&\quad + \mathbf{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \Delta_t \Delta_{S\cup\hat{H}_t,t} < 0\big\}\, \mathbf{1}\big\{4\theta_t \le \varepsilon\big\}\;. \qquad (17)
\end{aligned}$$

Next, we show that $\hat{B}_t$ can be replaced with B_t in the equation above. To do so, we use the fact that $\hat{B}_t$ appears only in terms that are multiplied by $\mathbf{1}\big\{4\theta_t \le \varepsilon\big\}$. Using the definition of $\hat{B}_t$, the fact that $|\hat{\Delta}_{j_t^\star,t}| \le |\hat{\Delta}_{\hat{j}_t,t}|$ and $|\Delta_{\hat{j}_t,t}| \le |\Delta_{j_t^\star,t}|$, together with the assumption $|\Delta_{j,t} - \hat{\Delta}_{j,t}| \le \theta$ for all $j \in [K]$ we get

$$\hat{B}_t \subseteq \{i \,:\, |\Delta_{j_t^\star,t}| - \tau - 4\theta_t \;\le\; |\Delta_{i,t}| \;\le\; |\Delta_{j_t^\star,t}| - \tau + 4\theta_t\}\;.$$

If $4\theta_t \le \varepsilon$ then the right-hand side above is a subset of $B_{\varepsilon,t}$, and therefore, under this condition, $\hat{B}_t \subseteq B_{\varepsilon,t}$. We conclude that $\hat{B}_t$ can be replaced by B_t in Equation (17), and

$$\begin{aligned}
Z_t &\le Z_t\, \mathbf{1}\big\{4\theta_t > \varepsilon\big\} + \mathbf{1}\big\{\exists S \subseteq B_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \le \theta_t\big\}\, \mathbf{1}\big\{4\theta_t \le \varepsilon\big\} \\
&\quad + \mathbf{1}\big\{\exists S \subseteq B_t \,:\, \Delta_t \Delta_{S\cup\hat{H}_t,t} < 0\big\}\, \mathbf{1}\big\{4\theta_t \le \varepsilon\big\} \\
&\le Z_t\, \mathbf{1}\big\{4\theta_t > \varepsilon\big\} + \mathbf{1}\big\{\exists S \subseteq B_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \le \varepsilon/4\big\} + \mathbf{1}\big\{\exists S \subseteq B_t \,:\, \Delta_t \Delta_{S\cup\hat{H}_t,t} < 0\big\}\;.
\end{aligned}$$

With the inequality above handy, we are now ready to upper-bound $Q_{T,\varepsilon}$. We have

$$\begin{aligned}
Q_{T,\varepsilon} &= \sum_{t=1}^{T} Z_t \, \mathbb{1}\big\{\forall S \subseteq B_{\varepsilon,t} \,:\, \Delta_t \Delta_{S\cup H_{\varepsilon,t},t} \geq 0,\ |\Delta_{S\cup H_{\varepsilon,t},t}| > \varepsilon\big\} \\
&\leq \sum_{t=1}^{T} Z_t \, \mathbb{1}\big\{4\theta_t > \varepsilon\big\} \\
&\quad + \underbrace{\mathbb{1}\big\{\exists S \subseteq B_t \,:\, |\hat{\Delta}_{S\cup \hat{H}_t,t}| \leq \varepsilon/4\big\}\, \mathbb{1}\big\{\forall S \subset B_{\varepsilon,t} \,:\, |\Delta_{S\cup H_{\varepsilon,t},t}| > \varepsilon\big\}}_{=0} \\
&\quad + \underbrace{\mathbb{1}\big\{\exists S \subseteq B_t \,:\, \Delta_t \Delta_{S\cup \hat{H}_t,t} < 0\big\}\, \mathbb{1}\big\{\forall S \subset B_{\varepsilon,t} \,:\, \Delta_t \Delta_{S\cup H_{\varepsilon,t},t} \geq 0\big\}}_{=0} \\
&\leq \frac{16}{\varepsilon^2} \sum_{t=1}^{T} Z_t \theta_t^2 \ .
\end{aligned}$$

Recall that $\theta_t^2 = \mathbf{x}_t^\top A_{t-1}^{-1} \mathbf{x}_t \big(1 + 4\sum_{i=1}^{t-1} Z_i r_i + 36\log(Kt/\delta)\big)$. Using Lemma 24 (i), we obtain $Q_{T,\varepsilon} \leq \frac{32}{\varepsilon^2} \sum_{t=1}^{T} Z_t r_t \Big(1 + 4\sum_{i=1}^{t-1} Z_i r_i + 36\log(Kt/\delta)\Big)$. The conclusion of the proof follows along the lines of the proof of Lemma 6.

Proof of Lemma 16. We first prove that $\hat{H}_t \subseteq C_t \subseteq \hat{C}_t$. If $j \in C_t$, then $|\Delta_{j,t}| \geq |\Delta_{j_t^\star,t}| - \tau \geq |\Delta_{\hat{j}_t,t}| - \tau$. Using the assumption that $|\Delta_{j,t} - \hat{\Delta}_{j,t}| \leq \theta_t$ and $|\Delta_{\hat{j}_t,t} - \hat{\Delta}_{\hat{j}_t,t}| \leq \theta_t$, we have that $|\hat{\Delta}_{j,t}| \geq |\hat{\Delta}_{\hat{j}_t,t}| - \tau - 2\theta_t$, and therefore $j \in \hat{C}_t$. Similarly, if $j \in \hat{H}_t$, then $|\hat{\Delta}_{j,t}| \geq |\hat{\Delta}_{\hat{j}_t,t}| - \tau + 2\theta_t \geq |\hat{\Delta}_{j_t^\star,t}| - \tau + 2\theta_t$. Using the assumption that $|\Delta_{j,t} - \hat{\Delta}_{j,t}| \leq \theta_t$ and $|\Delta_{j_t^\star,t} - \hat{\Delta}_{j_t^\star,t}| \leq \theta_t$, we get $|\Delta_{j,t}| \geq |\Delta_{j_t^\star,t}| - \tau$, and therefore $j \in C_t$.

Now assume that $Z_t = 0$. By definition, $\hat{\Delta}_t \hat{\Delta}_{S\cup\hat{H}_t,t} \geq 0$ and $|\hat{\Delta}_{S\cup\hat{H}_t,t}| > \theta$ for all $S \subseteq \hat{B}_t$, and particularly for $S = C_t \setminus \hat{H}_t$. Namely, $\hat{\Delta}_t \hat{\Delta}_{C_t,t} \geq 0$ and $|\hat{\Delta}_{C_t,t}| > \theta_t$. Once again using the assumption of the lemma, this time in conjunction with Jensen's inequality, we get that $(\Delta_t - \hat{\Delta}_{C_t,t})^2 \leq \theta_t^2$, which implies $\Delta_t \hat{\Delta}_{C_t,t} \geq \frac{1}{2}\Big(\hat{\Delta}_{C_t,t}^2 - \theta_t^2\Big)$. Plugging in $|\hat{\Delta}_{C_t,t}| > \theta_t$ gives $\Delta_t \hat{\Delta}_{C_t,t} > 0$ which, combined with $\hat{\Delta}_t \hat{\Delta}_{C_t,t} \geq 0$ gives $\Delta_t \hat{\Delta}_t \geq 0$. Overall we have shown that $Z_t = 0$ implies that $\Delta_t \hat{\Delta}_t \geq 0$. Therefore, $U_T = \sum_{t=1}^{T} \bar{Z}_t \, \mathbb{1}\big\{\Delta_t \hat{\Delta}_t < 0\big\} = 0$.

Proof of Lemma 19. First, note that, by the way Algorithm 2 is defined,

$$\begin{aligned}
Z_t &= \mathbb{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \hat{\Delta}_t \hat{\Delta}_{S\cup\hat{H}_t,t} < 0 \ \vee\ |\hat{\Delta}_{S\cup\hat{H}_t,t}| \leq \theta_{S\cup\hat{H}_t,t}\big\} \\
&= \mathbb{1}\big\{\exists S \subseteq \hat{B}_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \leq \theta_{S\cup\hat{H}_t,t}\big\} \\
&\quad + \mathbb{1}\big\{\forall S \subseteq \hat{B}_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| > \theta_{S\cup\hat{H}_t,t}\big\}\, \mathbb{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \hat{\Delta}_t \hat{\Delta}_{S\cup\hat{H}_t,t} < 0\big\} \\
&\leq \mathbb{1}\big\{\exists S \subseteq \hat{B}_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \leq \theta_{S\cup\hat{H}_t,t}\big\} \\
&\quad + \mathbb{1}\big\{|\hat{\Delta}_t| > \theta_{\hat{C}_t,t}\big\}\, \mathbb{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \hat{\Delta}_t \hat{\Delta}_{S\cup\hat{H}_t,t} < 0,\ |\hat{\Delta}_{S\cup\hat{H}_t,t}| > \theta_{S\cup\hat{H}_t,t}\big\}.
\end{aligned}$$

We focus on the second term on the right-hand side above. Using the assumption that $|\Delta_{j,t} - \hat{\Delta}_{j,t}| \leq \theta_{j,t}$ for all $j \in [K]$ together with Jensen's inequality, we have that $|\hat{\Delta}_t - \Delta_{\hat{C}_t,t}| \leq \theta_{\hat{C}_t,t}$ and $|\hat{\Delta}_{S\cup\hat{H}_t,t} - \Delta_{S\cup\hat{H}_t,t}| \leq \theta_{S\cup\hat{H}_t,t}$ for any S. Now, if S is such that $\hat{\Delta}_t \hat{\Delta}_{S\cup\hat{H}_t,t} < 0$, $|\hat{\Delta}_{S\cup\hat{H}_t,t}| > \theta_{S\cup\hat{H}_t,t}$, and $|\hat{\Delta}_t| > \theta_{\hat{C}_t,t}$, then it also holds that $\Delta_{\hat{C}_t,t} \Delta_{S\cup\hat{H}_t,t} < 0$. Moreover, if there exists $S \subseteq \hat{B}_t$ such that $\Delta_{\hat{C}_t,t} \Delta_{S\cup\hat{H}_t,t} < 0$ then either $\Delta_t \Delta_{S\cup\hat{H}_t,t} < 0$ or $\Delta_t \Delta_{\hat{C}_t,t} < 0$. Since $\hat{C}_t = \hat{H}_t \cup \hat{B}_t$ we have that

$$\mathbb{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \Delta_{\hat{C}_t,t} \Delta_{S\cup\hat{H}_t,t} < 0\big\} \ \leq\ \mathbb{1}\big\{\exists S \subseteq \hat{B}_t \,:\, \Delta_t \Delta_{S\cup\hat{H}_t,t} < 0\big\} \ .$$

Putting together, we can write

$$Z_t \;\leq\; \mathbf{1}\Big\{\exists S \subseteq \hat{B}_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \leq \theta_{S\cup\hat{H}_t,t}\Big\} + \mathbf{1}\Big\{\exists S \subseteq \hat{B}_t \,:\, \Delta_t \Delta_{S\cup\hat{H}_t,t} < 0\Big\} \;\; .$$

Using the above, we can decompose Z_t as follows

$$\begin{aligned}
Z_t \;&=\; Z_t\,\mathbf{1}\Big\{4\max_{j\in\hat{C}_t}\theta_{j,t} > \varepsilon\Big\} + Z_t\,\mathbf{1}\Big\{4\max_{j\in\hat{C}_t}\theta_{j,t} \leq \varepsilon\Big\} \\
&\leq\; Z_t\,\mathbf{1}\Big\{4\max_{j\in\hat{C}_t}\theta_{j,t} > \varepsilon\Big\} + \mathbf{1}\Big\{\exists S \subseteq \hat{B}_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \leq \theta_{S\cup\hat{H}_t,t}\Big\}\,\mathbf{1}\Big\{4\max_{j\in\hat{C}_t}\theta_{j,t} \leq \varepsilon\Big\} \\
&\quad+\; \mathbf{1}\Big\{\exists S \subseteq \hat{B}_t \,:\, \Delta_t\Delta_{S\cup\hat{H}_t,t} < 0\Big\}\,\mathbf{1}\Big\{4\max_{j\in\hat{C}_t}\theta_{j,t} \leq \varepsilon\Big\} \\
&\leq\; Z_t\,\mathbf{1}\Big\{4\max_{j\in\hat{C}_t}\theta_{j,t} > \varepsilon\Big\} + \mathbf{1}\Big\{\exists S \subseteq \hat{B}_t \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \leq \varepsilon/4\Big\}\,\mathbf{1}\Big\{4\max_{j\in\hat{C}_t}\theta_{j,t} \leq \varepsilon\Big\} \\
&\quad+\; \mathbf{1}\Big\{\exists S \subseteq \hat{B}_t \,:\, \Delta_t\Delta_{S\cup\hat{H}_t,t} < 0\Big\}\,\mathbf{1}\Big\{4\max_{j\in\hat{C}_t}\theta_{j,t} \leq \varepsilon\Big\} \;\; . \qquad (18)
\end{aligned}$$

where the last step is because $\max_{j\in\hat{C}_t}\theta_{j,t} \geq \theta_{S\cup\hat{H}_t,t}$. Next, we show that $\hat{B}_t$ can be replaced with B_t in the equation above. To do so, we use the fact that $\hat{B}_t$ appears only in terms that are multiplied by $\mathbf{1}\big\{4\max_{j\in\hat{C}_t}\theta_{j,t} \leq \varepsilon\big\}$. Using the definition of $\hat{B}_t$, the fact that $|\hat{\Delta}_{j_t^\star,t}| \leq |\hat{\Delta}_{\hat{j}_t,t}|$ and $|\Delta_{\hat{j}_t,t}| \leq |\Delta_{j_t^\star,t}|$, together with the assumption $|\Delta_{j,t} - \hat{\Delta}_{j,t}| \leq \theta_{j,t}$ for all $j \in [K]$ we get

$$\hat{B}_t \subseteq \left\{ i \,:\, |\Delta_{j_t^\star,t}| - \tau - 4\max_{j\in\hat{C}_t}\theta_{j,t} \;\leq\; |\Delta_{i,t}| \;\leq\; |\Delta_{j_t^\star,t}| - \tau + 4\theta_{\hat{C}_t,t} \right\} \;\; .$$

Hence, when $4\max_{j\in\hat{C}_t}\theta_{j,t} \leq \varepsilon$ we are guaranteed that the right-hand side above is a subset of $B_{\varepsilon,t}$, and therefore, under this condition, $\hat{B}_t \subseteq B_{\varepsilon,t}$. We conclude that $\hat{B}_t$ can be replaced by $B_{\varepsilon,t}$ in Equation (18), and so Z_t is upper bounded by

$$Z_t\,\mathbf{1}\Big\{4\max_{j\in\hat{C}_t}\theta_{j,t} > \varepsilon\Big\} + \mathbf{1}\Big\{\exists S \subseteq B_{\varepsilon,t} \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \leq \varepsilon/4\Big\} + \;\mathbf{1}\Big\{\exists S \subseteq B_{\varepsilon,t} \,:\, \Delta_t\Delta_{S\cup\hat{H}_t,t} < 0\Big\} \;\; .$$

With the inequality above handy, we are now ready to upper-bound $Q_{T,\varepsilon}$. We have

$$\begin{aligned}
Q_{T,\varepsilon} &= \sum_{t=1}^{T} Z_t \, \mathbb{1}\left\{\forall S \subseteq B_{\varepsilon,t} \,:\, \Delta_t \Delta_{S\cup H_{\varepsilon,t},t} \geq 0,\ |\Delta_{S\cup H_{\varepsilon,t},t}| > \varepsilon\right\} \\
&\leq \sum_{t=1}^{T} Z_t \, \mathbb{1}\left\{4 \max_{j\in\hat{C}_t} \theta_{j,t} > \varepsilon\right\} \\
&\quad + \underbrace{\mathbb{1}\left\{\exists S \subseteq B_{\varepsilon,t} \,:\, |\hat{\Delta}_{S\cup\hat{H}_t,t}| \leq \varepsilon/4\right\} \mathbb{1}\left\{\forall S \subset B_{\varepsilon,t} \,:\, |\Delta_{S\cup H_{\varepsilon,t},t}| > \varepsilon\right\}}_{=0} \\
&\quad + \underbrace{\mathbb{1}\left\{\exists S \subseteq B_{\varepsilon,t} \,:\, \Delta_t \Delta_{S\cup\hat{H}_t,t} < 0\right\} \mathbb{1}\left\{\forall S \subset B_{\varepsilon,t} \,:\, \Delta_t \Delta_{S\cup H_{\varepsilon,t},t} \geq 0\right\}}_{=0} \\
&\leq \frac{16}{\varepsilon^2} \sum_{t=1}^{T} Z_t \max_{j\in\hat{C}_t} \theta_{j,t}^2 \\
&\leq \frac{16 \sum_{t=1}^{T} Z_t \sum_{j\in\hat{C}_t} \theta_{j,t}^2}{\varepsilon^2} \\
&= \frac{16 \sum_{j\in[K]} \sum_{t=1}^{T} Z_t \, \mathbb{1}\left\{j \in \hat{C}_t\right\} \theta_{j,t}^2}{\varepsilon^2} \\
&= \frac{16 \sum_{j\in[K]} \sum_{t=1}^{T} Z_t \, \mathbb{1}\left\{j \in \hat{C}_t\right\} \mathbf{x}_t^\top A_{j,t-1}^{-1} \mathbf{x}_t \left(1 + 4\sum_{i=1}^{t-1} Z_i r_{j,i} + 36\log(Kt/\delta)\right)}{\varepsilon^2} \\
&= \frac{16 \sum_{j\in[K]} \sum_{t=1}^{T} Z_t r_{j,t} \left(1 + 4\sum_{i=1}^{t-1} Z_i r_{j,i} + 36\log(Kt/\delta)\right)}{\varepsilon^2} \,.
\end{aligned}$$

Now, proceeding along the same lines as in the proof of Lemma 6 (which in turn mainly relies on Lemma 24) we conclude that,

$$Q_{T,\varepsilon} \leq \frac{16 \sum_{j\in[K]} \left((1 + 36\log(KT/\delta)) \log|A_{j,T}| + 4\log^2|A_{j,T}|\right)}{\varepsilon^2} = O\left(\frac{Kd^2 \log^2(KT/\delta)}{\varepsilon^2}\right) .$$

This concludes the proof.

Proof of Lemma 20. The proof proceeds in the same way as the proof of Lemma 16. We first prove that $\hat{H}_t \subseteq C_t \subseteq \hat{C}_t$. If $j \in C_t$, then $|\Delta_{j,t}| \geq |\Delta_{j_t^\star,t}| - \tau \geq |\Delta_{\hat{j}_t,t}| - \tau$. Using the assumption that $|\Delta_{j,t} - \hat{\Delta}_{j,t}| \leq \theta_{j,t}$ and $|\Delta_{\hat{j}_t,t} - \hat{\Delta}_{\hat{j}_t,t}| \leq \theta_{\hat{j}_t,t}$, we have that $|\hat{\Delta}_{j,t}| \geq |\hat{\Delta}_{\hat{j}_t,t}| - \tau - \theta_{j,t} - \theta_{\hat{j}_t,t}$, and therefore $j \in \hat{C}_t$. Similarly, if $j \in \hat{H}_t$, then

$$|\hat{\Delta}_{j,t}| \geq |\hat{\Delta}_{\hat{j}_t,t}| - \tau + \theta_{j,t} + \max_{j\in\hat{C}_t} \geq |\hat{\Delta}_{j_t^\star,t}| - \tau + \theta_{j,t} + \max_{j\in\hat{C}_t} .$$

Using the assumption that $|\Delta_{j,t} - \hat{\Delta}_{j,t}| \leq \theta_{j,t}$ and $|\Delta_{j_t^\star,t} - \hat{\Delta}_{j_t^\star,t}| \leq \theta_{j_t^\star,t} \leq \max_{j\in\hat{C}_t} \theta_{j,t}$, we get $|\Delta_{j,t}| \geq |\Delta_{j_t^\star,t}| - \tau$, and therefore $j \in C_t$.

Now assume that $Z_t = 0$. By definition, $\hat{\Delta}_t \hat{\Delta}_{S\cup\hat{H}_t,t} \geq 0$ and $|\hat{\Delta}_{S\cup\hat{H}_t,t}| > \theta_{S\cup\hat{H}_t,t}$ for all $S \subseteq \hat{B}_t$, and particularly for $S = C_t \setminus \hat{H}_t$. Namely, $\hat{\Delta}_t \hat{\Delta}_{C_t,t} \geq 0$ and $|\hat{\Delta}_{C_t,t}| > \theta_{\hat{C}_t,t}$. Once again using the assumption of the lemma, this time in conjunction with Jensen's inequality, we get that $(\Delta_t - \hat{\Delta}_{C_t,t})^2 \leq \theta_t^2$, which implies $\Delta_t \hat{\Delta}_{C_t,t} \geq \frac{1}{2}\left(\hat{\Delta}_{C_t,t}^2 - \theta_{\hat{C}_t,t}^2\right)$. Plugging in $|\hat{\Delta}_{C_t,t}| > \theta_{\hat{C}_t,t}$ gives $\Delta_t \hat{\Delta}_{C_t,t} > 0$ which, combined

with $\hat{\Delta}_t \hat{\Delta}_{C_t,t} \geq 0$, gives $\Delta_t \hat{\Delta}_t \geq 0$. Overall we have shown that $Z_t = 0$ implies that $\Delta_t \hat{\Delta}_t \geq 0$. Therefore, $U_T \;=\; \sum_{t=1}^{T} \bar{Z}_t \, \mathbb{1}\{\Delta_t \hat{\Delta}_t < 0\} \;=\; 0$.

Proof of Lemma 18. We start just as in the proof of Lemma 14 and get,

$$\begin{aligned} Z_t \leq \;& \mathbb{1}\{|\Delta_t| \leq \varepsilon\} \;+\; Z_t \, \mathbb{1}\{|\Delta_t| > \varepsilon\} \, \mathbb{1}\{\exists S \subseteq B_{\varepsilon,t} \;:\; \Delta_t \Delta_{S \cup H_{\varepsilon,t}} < 0 \;\vee\; |\Delta_{S \cup H_{\varepsilon,t}}| \leq \varepsilon\} \\ & + Z_t \, \mathbb{1}\{\forall S \subseteq B_{\varepsilon,t} \;:\; \Delta_t \Delta_{S \cup H_{\varepsilon,t}} \geq 0, \; |\Delta_{S \cup H_{\varepsilon,t}}| > \varepsilon\} \;\; . \end{aligned}$$

Hence,

$$\begin{aligned} N_T = \;& \sum_{t=1}^{T} |\hat{C}_t| Z_t \\ \leq \;& \sum_{t=1}^{T} |\hat{C}_t| \, \mathbb{1}\{|\Delta_t| \leq \varepsilon\} \\ & + \sum_{t=1}^{T} |\hat{C}_t| Z_t \, \mathbb{1}\{|\Delta_t| > \varepsilon\} \, \mathbb{1}\{\exists S \subseteq B_{\varepsilon,t} \;:\; \Delta_t \Delta_{S \cup H_{\varepsilon,t}} < 0 \;\vee\; |\Delta_{S \cup H_{\varepsilon,t}}| \leq \varepsilon\} \\ & + \sum_{t=1}^{T} |\hat{C}_t| Z_t \, \mathbb{1}\{\forall S \subseteq B_{\varepsilon,t} \;:\; \Delta_t \Delta_{S \cup H_{\varepsilon,t}} \geq 0, \; |\Delta_{S \cup H_{\varepsilon,t}}| > \varepsilon\} \\ \leq \;& K \sum_{t=1}^{T} \mathbb{1}\{|\Delta_t| \leq \varepsilon\} \\ & + K \sum_{t=1}^{T} \frac{|\hat{C}_t|}{K} Z_t \, \mathbb{1}\{|\Delta_t| > \varepsilon\} \, \mathbb{1}\{\exists S \subseteq B_{\varepsilon,t} \;:\; \Delta_t \Delta_{S \cup H_{\varepsilon,t}} < 0 \;\vee\; |\Delta_{S \cup H_{\varepsilon,t}}| \leq \varepsilon\} \\ & + K \sum_{t=1}^{T} Z_t \, \mathbb{1}\{\forall S \subseteq B_{\varepsilon,t} \;:\; \Delta_t \Delta_{S \cup H_{\varepsilon,t}} \geq 0, \; |\Delta_{S \cup H_{\varepsilon,t}}| > \varepsilon\} \\ = \;& K T_\varepsilon \\ & + K \sum_{t=1}^{T} \tfrac{|\hat{C}_t| Z_t}{K} \, \mathbb{1}\{|\Delta_t| > \varepsilon\} \, \mathbb{1}\{\exists S \subseteq B_{\varepsilon,t} : \Delta_t \Delta_{S \cup H_{\varepsilon,t}} < 0 \vee |\Delta_{S \cup H_{\varepsilon,t}}| \leq \varepsilon\} \\ & + K Q_{T,\varepsilon} \; . \end{aligned} \tag{19}$$

Now we note that by definition of $\hat{C}_t$, if $j \in \hat{C}_t$ then

$$|\hat{\Delta}_{j,t}| \geq |\hat{\Delta}_{\hat{j}_t,t}| - \tau - \theta_{j,t} - \theta_{\hat{j}_t,t} \geq |\hat{\Delta}_{j_t^\star,t}| - \tau - \theta_{j,t} - \theta_{\hat{j}_t,t} \; .$$

Combined with our assumption that $(\Delta_{j,t} - \hat{\Delta}_{j,t})^2 \leq \theta_{j,t}^2$ holds for all $j \in [K]$ this implies

$$|\Delta_{j,t}| \geq |\Delta_{j_t^\star,t}| - \tau - 2\theta_{j,t} - \theta_{\hat{j}_t,t} - \theta_{j_t^\star,t} \; . \tag{20}$$

On the other hand, by definition of $j_t^\star$, we also have $|\Delta_{j_t^\star,t}| \geq |\Delta_{\hat{j}_t,t}|$ and, owing to our assumption, $|\hat{\Delta}_{j_t^\star,t}| \geq |\hat{\Delta}_{\hat{j}_t,t}| - \theta_{j_t^\star,t} - \theta_{\hat{j}_t,t}$. Hence we see that $j_t^\star \in \hat{C}_t$. Using this in Equation (20) gives, for any $j \in \hat{C}_t$,

$$|\Delta_{j,t}| \geq |\Delta_{j_t^\star,t}| - \tau - 4 \max_{j \in \hat{C}_t} \theta_{j,t} \; .$$

Thus we see that as long as $4 \max_{j \in \hat{C}_t} \theta_{j,t} \leq \varepsilon$, we have $\hat{C}_t \subset H_{\varepsilon,t} \cup B_{\varepsilon,t}$.

We now use this in Equation (19). We obtain

$$\begin{aligned}
N_T \leq & K T_\varepsilon + K Q_{T,\varepsilon} \\
& + K \left(\sum_{t=1}^{T} \frac{|\hat{C}_t| Z_t}{K} \mathbb{1}\left\{ 4 \max_{j \in \hat{C}_t} \theta_{j,t} \leq \varepsilon \right\} \mathbb{1}\left\{ |\Delta_t| > \varepsilon \right\} \times \right. \\
& \qquad \left. \times \mathbb{1}\left\{ \exists S \subseteq B_{\varepsilon,t} : \Delta_t \Delta_{S \cup H_{\varepsilon,t}} < 0 \vee |\Delta_{S \cup H_{\varepsilon,t}}| \leq \varepsilon \right\} \right) \\
& + K \left(\sum_{t=1}^{T} \frac{|\hat{C}_t| Z_t}{K} \mathbb{1}\left\{ 4 \max_{j \in \hat{C}_t} \theta_{j,t} > \varepsilon \right\} \mathbb{1}\left\{ |\Delta_t| > \varepsilon \right\} \times \right. \\
& \qquad \left. \times \mathbb{1}\left\{ \exists S \subseteq B_{\varepsilon,t} : \Delta_t \Delta_{S \cup H_{\varepsilon,t}} < 0 \vee |\Delta_{S \cup H_{\varepsilon,t}}| \leq \varepsilon \right\} \right) \\
\leq & \; K T_\varepsilon + K Q_{T,\varepsilon} \\
& + K \left(\sum_{t=1}^{T} \frac{|B_{\varepsilon,t} \cup H_{\varepsilon,t}|}{K} \mathbb{1}\left\{ |\Delta_t| > \varepsilon \right\} \mathbb{1}\left\{ \exists S \subseteq B_{\varepsilon,t} : \Delta_t \Delta_{S \cup H_{\varepsilon,t}} < 0 \vee |\Delta_{S \cup H_{\varepsilon,t}}| \leq \varepsilon \right\} \right) \\
& + K \sum_{t=1}^{T} Z_t \mathbb{1}\left\{ 4 \max_{j \in \hat{C}_t} \theta_{j,t} > \varepsilon \right\} \\
= & \; K \left(T_\varepsilon + T''_\varepsilon + Q_{T,\varepsilon} + \sum_{t=1}^{T} Z_t \mathbb{1}\left\{ 4 \max_{j \in \hat{C}_t} \theta_{j,t} > \varepsilon \right\} \right) . \qquad (21)
\end{aligned}$$

In order to bound last term, we notice that

$$\begin{aligned}
& \sum_{t=1}^{T} Z_t \mathbb{1}\left\{ 4 \max_{j \in \hat{C}_t} \theta_{j,t} > \varepsilon \right\} \\
& \quad \leq \frac{16 \sum_{t=1}^{T} Z_t \max_{j \in \hat{C}_t} \theta_{j,t}^2}{\varepsilon^2} \\
& \quad \leq \frac{16 \sum_{t=1}^{T} Z_t \sum_{j \in \hat{C}_t} \theta_{j,t}^2}{\varepsilon^2} \\
& \quad = \frac{16 \sum_{j \in [K]} \sum_{t=1}^{T} Z_t \mathbb{1}\left\{ j \in \hat{C}_t \right\} \theta_{j,t}^2}{\varepsilon^2} \\
& \quad = \frac{16 \sum_{j \in [K]} \sum_{t=1}^{T} Z_t \mathbb{1}\left\{ j \in \hat{C}_t \right\} \mathbf{x}_t^\top A_{j,t-1}^{-1} \mathbf{x}_t \left(1 + 4 \sum_{i=1}^{t-1} Z_i r_{j,i} + 36 \log(Kt/\delta) \right)}{\varepsilon^2} \\
& \quad = \frac{16 \sum_{j \in [K]} \sum_{t=1}^{T} Z_t r_{j,t} \left(1 + 4 \sum_{i=1}^{t-1} Z_i r_{j,i} + 36 \log(Kt/\delta) \right)}{\varepsilon^2} .
\end{aligned}$$

Now, proceeding along the lines of the proof of Lemma 6 (which in turn mainly relies on Lemma 24), we obtain

$$\begin{aligned}
\sum_{t=1}^{T} Z_t \mathbb{1}\left\{ 4 \max_{j \in \hat{C}_t} \theta_{j,t} > \varepsilon \right\} & \leq \frac{16 \sum_{j \in [K]} \left((1 + 36 \log(KT/\delta)) \log |A_{j,T}| + 4 \log^2 |A_{j,T}| \right)}{\varepsilon^2} \\
& = O\left(\frac{K d^2 \log^2(KT/\delta)}{\varepsilon^2} \right) .
\end{aligned}$$

Plugging these back into Equation (21) and applying the bound on $Q_{T,\varepsilon}$ from Lemma 19 concludes the proof.

References

B. D. Argall, B. Browning, and M. Veloso. Automatic weight learning for multiple data sources when learning from demonstration. In *Proc. of the 2009 IEEE International Conference on Robotics and Automation*, pages 226–231, 2009.

K. S. Azoury and M. K. Warmuth. Relative loss bounds for online density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

F. Bach. Active learning for misspecified generalized linear models. Technical Report N15/06/MM, Ecole des mines de Paris, June 2006.

M. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proc. of the 23th International Conference on Machine Learning*, pages 65–72, 2006.

M. Balcan, A. Broder, and T. Zhang. Margin-based active learning. In *Proc. of the 20th Annual Conference on Learning Theory*, pages 35–50, 2007.

M. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *Proc. of the 21th Annual Conference on Learning Theory*, pages 45–56, 2008.

A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems 24*, pages 199–207, 2010.

A. Beygelzimer, D. Hsu, N. Karampatziakis, J. Langford, and T. Zhang. Efficient active learning. In *ICML 2011 Workshop on On-line Trading of Exploration and Exploitation*, 2011.

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

G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear classification and selective sampling under low noise conditions. In *Advances in Neural Information Processing Systems 21*, pages 249–256, 2009.

N. Cesa-Bianchi and C. Gentile. Improved risk tail bounds for on-line algorithms. *IEEE Transactions on Information Theory*, 54(1):386–390, 2008.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. Learning probabilistic linear-threshold classifiers via selective sampling. In *Proc. of the 16th Annual Conference on Learning Theory*, pages 373–387, 2003.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. A second-order Perceptron algorithm. *SIAM Journal on Computing*, 43(3):640–668, 2005a.

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005b.

N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Worst-case analysis of selective sampling for linear classification. *Journal of Machine Learning Research*, 7:1025–1230, 2006.

N. Cesa-Bianchi, C. Gentile, and F. Orabona. Robust bounds for classification via selective sampling. In *Proc. of the 26th International Conference on Machine Learning*, pages 121–128, 2009.

S. Chen, J. Zhang, G. Chen, and C Zhang. What if the irresponsible teachers are dominating? A method of training on samples and clustering on teachers. In *Proc. of the Twenty-fourth AAAI Conference on Artificial Intelligence*, pages 419–424, 2010.

R. Cohn, L. Atlas, and R. Ladner. Training connectionist networks with queries and selective sampling. In *Advances in Neural Information Processing Systems 2*, pages 566–573, 1990.

K. Crammer and C. Gentile. Multiclass classification with bandit feedback using adaptive regularization. In *Proc. of the 28th International Conference on Machine Learning*, pages 273–280, 2011.

K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9:1757–1774, 2008.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Proc. of the 21th Annual Conference on Learning Theory*, pages 355–366, 2008.

S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proc. of the 18th Annual Conference on Learning Theory*, pages 249–263, 2005.

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems 21*, pages 353–360, 2008.

A.P. Dawid and A.M. Skeene. Maximum likelihood estimation of observed error-rates using the em algorithm. *Applied statistics*, 28:20–28, 1979.

O. Dekel and O. Shamir. Vox populi: Collecting high-quality labels from a crowd. In *Proc. of the 22nd Annual Conference on Learning Theory*, 2009a.

O. Dekel and O. Shamir. Good learners for evil teachers. In *Proc. of the Twenty-Sixth International Conference on Machine Learning*, pages 216–223, 2009b.

P. Domnez. *Proactive Learning: Towards Learning with Multiple Imperfect Predictors*. PhD thesis, Carnegie Mellon University, 2010.

P. Donmez and J. G. Carbonell. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proc. of the 17th ACM Conference on Information and Knowledge Management*, pages 619–628, 2008.

P. Donmez, J.G. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proc. of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2009.

P. Donmez, J.G. Carbonell, and J. Schneider. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In *Proc. of the SIAM International Conference on Data Mining*, pages 826–837, 2010.

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

P. Groot, A. Birlutiu, and T. Heskes. Learning from multiple annotators with Gaussian processes. In *Proc. of the 21st International Conference on Artificial Neural Networks*, pages 159–164, 2011.

S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proc. of the 24th International Conference on Machine Learning*, pages 353–360, 2007.

S. Hanneke. Adaptive rates of convergence in active learning. In *Proc. of the 22th Annual Conference on Learning Theory*, 2009.

E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 2:169–192, 2007.

D. Helmbold and S. Panizza. Some label efficient learning results. In *Proc. of the 10th Conference of Computationa Learning Theory*, pages 218–230, 1997.

A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

S.L. Hui and X.H. Zhou. Evaluation of diagnostic tests without a gold standard. *Statistical Methods in Medical Research*, 7:354–370, 1998.

S. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2008.

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

T. L. Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, pages 154–166, 1982.

L. Li, M. Littman, and T. Walsh. Knows what it knows: a framework for self-aware learning. In *Proc. of the 25th International Conference on Machine Learning*, pages 568–575, 2008.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems 21*, pages 1041–1048, 2009a.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the Renyi divergence. In *Proc. of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 367–374, 2009b.

P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. Economical active feature-value acquisition through expected utility estimation. In *KDD Workshop on Utility-based data mining*, 2005.

F. Orabona and N. Cesa-Bianchi. Better algorithms for selective sampling. In *Proc. of the 28th International Conference on Machine Learning*, pages 433–440, 2011.

V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning*, 11:1297–1322, 2010.

V.S. Sheng, F. Provost, and P.G. Ipeirotis. Get another label? Improving data quality and data mining using multiple noisy labelers. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622, 2008.

P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of venus images. In *Advances in neural information processing systems*, pages 1085–1092, 1995.

R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

D.J. Spiegelhalter and P. Stovin. An analysis of repeated biopsies following cardiac transplantation. *Statistics in Medicine*, 2(1):33–40, 1983.

A. Strehl and M. Littman. Online linear regression and its application to model-based reinforcement learning. In *Advances in Neural Information Processing Systems 20*, pages 1417–1424, 2008.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1): 135–166, 2004.

V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

Y. Yan, R. Rosales, L. Bogoni, G. Fung, L. Moy, M. Schmidt, and J.G. Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proc. of the 13th International Conference on Artificial Intelligence and Statistics*, pages 932–939, 2010.

Y. Yan, R. Rosales, G. Fung, and J. Dy. Active learning from crowds. In *Proc. of the 28th International Conference on Machine Learning*, pages 1161–1168, 2011.

L. Yang and J. Carbonell. Cost complexity of proactive learning via a reduction to realizable active learning. Technical Report CMU-ML-09-113, Carnegie Mellon University, 2009a.

L. Yang and J. Carbonell. Adaptive proactive learning with cost-reliability tradeoff. Technical Report CMU-ML-09-114, Carnegie Mellon University, 2009b.