

Active Learning in the Non-realizable Case

Matti Kääriäinen

Department of Computer Science
University of Helsinki
`matti.kaariainen@cs.helsinki.fi`

Abstract. Most of the existing active learning algorithms are based on the realizability assumption: The learner’s hypothesis class is assumed to contain a target function that perfectly classifies all training and test examples. This assumption can hardly ever be justified in practice. In this paper, we study how relaxing the realizability assumption affects the sample complexity of active learning. First, we extend existing results on query learning to show that any active learning algorithm for the realizable case can be transformed to tolerate random bounded rate class noise. Thus, bounded rate class noise adds little extra complications to active learning, and in particular exponential label complexity savings over passive learning are still possible. However, it is questionable whether this noise model is any more realistic in practice than assuming no noise at all.

Our second result shows that if we move to the truly non-realizable model of statistical learning theory, then the label complexity of active learning has the same dependence $\Omega(1/\epsilon^2)$ on the accuracy parameter ϵ as the passive learning label complexity. More specifically, we show that under the assumption that the best classifier in the learner’s hypothesis class has generalization error at most $\beta > 0$, the label complexity of active learning is $\Omega(\beta^2/\epsilon^2 \log(1/\delta))$, where the accuracy parameter ϵ measures how close to optimal within the hypothesis class the active learner has to get and δ is the confidence parameter. The implication of this lower bound is that exponential savings should not be expected in realistic models of active learning, and thus the label complexity goals in active learning should be refined.

1 Introduction

In standard passive (semi)supervised learning, the labeled (sub)sample of training examples is generated randomly by an unknown distribution defining the learning problem. In contrast, an active learner has some control over which examples are to be labeled during the training phase. Depending on the specifics of the learning model, the examples to be labeled can be selected from a pool of unlabeled data, filtered online from a stream of unlabeled examples, or synthesized by the learner. The motivation for active learning is that label information is often expensive, and so training costs can potentially be reduced significantly by concentrating the labeling efforts on examples that the learner considers useful.

This hope is supported by both theoretical and practical evidence: There exist active learning algorithms that in certain restricted settings give provably exponential savings in label complexity [1–4], and also a variety of heuristic methods that at least sometimes give significant label complexity savings in practice (see, e.g., [5] and the references therein).

Unfortunately, there is a huge gap between the theory and the practice of active learning, even without considering computational complexity issues. The theoretical methods rely on unrealistic assumptions that render them (or at least their analysis) inapplicable to real life learning problems, while the practically motivated heuristics have no associated theoretical guarantees and indeed often fail miserably. One of the most unrealistic assumptions common to virtually all theoretical work in active learning is the realizability (or PAC) assumption, i.e., the assumption that the correct labeling is given by a target function belonging to a known hypothesis class F . The realizability assumption is never true in practice — at least we are aware of no real world problem in which it could be justified — and seems to lead to fragile learning algorithms that may and often do completely break down when the problem turns out to be non-realizable. Thus, relaxing the realizability assumption is a necessary first step in making the theory of active learning relevant to practice.

Many relaxations to the realizability assumption have been studied in passive learning, but it is not at all clear which of them leads to the best model for active learning. If the assumptions are relaxed too little, the theory may remain inapplicable. On the other hand, if no restrictions on noise are imposed, learning becomes impossible. In this paper, we try to chart where exactly the fruitful regime for active learning resides on the range between full realizability and arbitrary adversarial noise. To this end, we study two relaxations to the realizability assumption, both adapted to active learning from passive learning.

First, we show that in the model of bounded rate class noise [6], active learning is essentially as easy as in the realizable case, provided that the noise is non-persistent, i.e., each label query is corrupted independently at random. The key idea is to cancel out the noise by repeating each query a sufficient number of times and using the majority of the answers as a proxy for the true label. This way, any active learning algorithm for the realizable case can be transformed to tolerate bounded rate class noise with the cost of increasing the label complexity by a factor that has optimal dependence on the noise rate and logarithmic dependence on the original label complexity in the realizable case. Applying the transformation to an optimal algorithm for the realizable case yields a close to optimal algorithm for the bounded rate class noise case, so there is no need to design active learning algorithms for the bounded rate class noise model separately. Our strategy of repeated queries is a simplification of a similar strategy independently proposed in the query learning context [7], but unlike the earlier solution, our adaptive sampling based strategy requires no prior knowledge on the noise rate nor a separate step for estimating an upper bound for it.

The noise cancelling transformation makes the bounded rate class noise model look quite suspicious: In theory, the strategy of repeated queries is close to optimal, yet in practice it would most likely fail. The reason for the likely failure is of course that in reality non-realizability is rarely caused by random noise alone. Instead, non-realizability typically arises also from the aspects of the learning problem that are not sufficiently well understood or well-behaved to be modelled accurately. In case none of the models fits the learning problem perfectly, justifying assumptions on how exactly the modelling assumptions (say, linear separability in feature space) are violated is hard. To cope with such deviations, a more general model for non-realizability is needed.

In the model of statistical learning theory [8] the data is assumed to be generated iid by some unknown distribution that defines the learning problem, but the relationship between the objects and the labels is only implicitly assumed to be approximately representable by some classifier in the learner's hypothesis class. This model has been very fruitful in passive learning: Not only has it resulted in a body of new interesting theory, but also in many successful algorithms (e.g., soft-margin SVMs) that indeed seem to handle the types of non-realizability encountered in practice quite nicely. We believe similar success is possible also in active learning. Firstly, the empirical observation that the statistical learning theory model fits nicely into many real learning problems (in the sense that the algorithms developed within the model work well in practice) remains true in the active learning case, as the learning problems are more or less the same. Secondly, recent results show that the model is benevolent enough to make non-trivial positive results in active learning possible. In particular, [4] presents an algorithm for the statistical learning theory framework that exhibits exponential label complexity savings, but only until a certain accuracy threshold is reached. Whether exponential savings are possible on all accuracy levels, i.e., whether active learning can improve the learning rate in the statistical learning theory framework, is an open question.

The main contribution of this paper is to show that the kind of exponential label complexity savings that can sometimes be obtained in the realizable case¹ are not possible if true non-realizability as in the statistical learning theory model is allowed. We show that if the realizability assumption is relaxed by assuming that the learning problem is such that the generalization error of the best classifier in the hypothesis class F is at most β , then the expected label complexity on some such problems has to be at least $\Omega(\beta^2/\epsilon^2)$ to guarantee that the generalization error of the hypothesis output by the active learner is within ϵ of optimal in F . We show this even in the noise-free case, i.e., when the labels are fully determined by a target function (possibly outside F), thus showing that the lower bound arises from non-realizability per se and not random noise as is the case with the lower bound for learning with membership queries in [9]. Also, the lower bound remains true even when the unlabeled examples are uniformly distributed.

If only the dependence on ϵ is concerned, the lower bound matches the upper bound $O(1/\epsilon^2)$ for passive learning. Thus, allowing true non-realizability makes

¹ And in the bounded rate class noise case by our noise-cancelling transformation.

active learning require the same order of labeled examples as passive learning does, which is in huge contrast to the realizable case in which exponential savings are sometimes known to be possible. This justifies our choice of not studying the more adversarial models of malicious noise that have been studied in passive learning, since already allowing arbitrary non-malicious errors seems to kill most of the potential of active learning. Our lower bound matches the above mentioned upper bound proved in [4] that shows that active learning can drop the label complexity exponentially even in the truly non-realizable case when the target accuracy ϵ is large in comparison to β . Combined, the results show that active learning can indeed help exponentially much in the initial phase of learning, after which the gains deteriorate and the speed of learning drops to match that of passive learning. This prediction is well in line with the empirical observations that active learning heuristics tend to initially clearly outperform passive learning algorithms, but become less useful or even harmful as the learning progresses.

In contrast to the recent label complexity lower bounds of $\Omega(1/\epsilon)$ for active learning in the realizable case [3], our lower bound does not depend on special properties of F or the data distribution, but applies whenever F contains at least two classifiers that sometimes agree and that disagree on an unbounded set. Also, our lower bound is better by a factor of $1/\epsilon$, which is to be expected due to non-realizability.

The rest of the paper is organized as follows. In Section 2, we introduce the active learning framework used in this paper. Section 3 is devoted to our positive result, showing how bounded rate class noise can be dealt with. Then, we move on to the more realistic full non-realizability assumption and prove our lower bound for that case in Section 4. Finally, the conclusions are presented in Section 5.

2 Learning Model

Throughout the following, we assume that the learning problem is modeled by a probability distribution P on (object, label) pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, whose marginal on \mathcal{X} is denoted by P_X . The goal of the active learning algorithm is to find a classifier f with small generalization error $\epsilon(f) = \mathbb{P}(f(X) \neq Y)$ with respect to this distribution. We compare the generalization error of the learned classifier to $\min_{f \in F} \epsilon(f)$. Here, the hypothesis class F is used only as a comparison class, so the learner is not restricted to select its classifier from F . The realizable case is the special case where $\mathbb{P}(Y = f(X)) = 1$ for some $f \in F$.

What differentiates active learning from passive learning is that we assume that the active learner can choose the examples whose labels it wants to query. The amount and type of control in choosing the queries varies among different active learning models and ranges from complete freedom (query learning with membership queries), to selecting the query points from a pool of unlabeled data (pool-based active learning), to deciding online which queries to make while observing a stream of unlabeled data (filtering based active learning). Out of

these variants, different flavors of the last two are considered the central active learning models.

Each variant of active learning has its own motivation, and since we do not have to, we will not commit to any one of them. Instead, we formulate our results so that they apply to a template active learning algorithm that is flexible enough to cover all the above mentioned active learning models simultaneously. The template is presented in Figure 1. Here, $\text{Teacher}(x)$ denotes the label oracle, which according to our assumption of the data being iid samples from P implies that $\text{Teacher}(x) \sim P(Y|X = x)$, and that the answers of the teacher are independent given the query points.

```

ActiveLearn( $\epsilon, \delta$ )
 $U$  = pool of unlabeled data sampled iid from  $P_X$ 
do
  choose query point  $x \in U$ 
  query  $y = \text{Teacher}(x)$ 
  add more points to  $U$  by sampling from  $P_X$ 
while (!stopping_condition)
output  $f \in F$ 

```

Fig. 1. Template for active learning algorithms.

The template defines how the active learner can access P . As long as P is not accessed except as seen in Figure 1, the active learner can be completely arbitrary and possibly randomized. The gray parts are optional, and the inclusion or exclusion thereof leads to different restricted models of active learning. More specifically, including all the gray parts corresponds to the general active learning algorithm in [3], including the constraint on query points being chosen from U which itself is not updated corresponds to pool-based active learning, and the case in which arbitrary label queries are allowed corresponds to query learning.

An active learning algorithm is defined to be (ϵ, δ) -successful with respect to a class of learning problems \mathcal{P} and a hypothesis class F if, for all learning problems $P \in \mathcal{P}$, the generalization error of the classifier f output by the active learner is with probability at least $1 - \delta$ (over the examples and the randomness in the learner) within ϵ of the generalization error of the best classifier in F . The key quantities of interest to us are the random number $n(\epsilon, \delta)$ of queries to Teacher , also known as the active learning label complexity, and the number of unlabeled examples $m(\epsilon, \delta)$ the active learner samples from P_X . Even though labeled examples are typically assumed to be far more expensive than unlabeled examples, the latter cannot be assumed to be completely free (since already processing them costs something). Thus, the goal is to be successful with as small (expected) $n(\epsilon, \delta)$ as possible, while keeping the (expectation of) $m(\epsilon, \delta)$ non-astronomical.

Of course, the difficulty of active learning depends on what we assume of the underlying task P and also on what we compare our performance to. In our definition, these are controlled by the choice of \mathcal{P} and the comparison class F . One extreme is the realizability assumption that corresponds to the assumption

that

$$\mathcal{P}_F = \{P \mid \exists f \in F : P(f(X) = Y) = 1\},$$

and choosing the comparison class to be the same F in which the target is assumed to reside. As already mentioned, in this special case exponential savings are possible in case F is the class of threshold functions in one dimension [1]. Also, if F is the class of linear separators going through the origin in R^d , and in addition to realizability we assume that the distribution P_X is uniform on the unit sphere of R^d , successful active learning is possible with $n = O(\log(1/\epsilon))$ label queries and $m = O(1/\epsilon)$ unlabeled examples, whereas the same task requires $n = \Omega(1/\epsilon)$ labeled examples in passive learning. Here, the dependence on all other parameters like d and δ has been abstracted away, so only the rate as a function of the accuracy parameter ϵ is considered. For algorithms achieving the above mentioned rates, see [2, 3].

The above cited results for the realizable case show that active learning can in some special cases give exponential savings, and this has lead some researchers to believe that such savings might be possible also for other function classes, without assumptions on P_X , and also without the realizability assumption. However, there is little concrete evidence supporting such beliefs.

3 Positive Result

Let us first replace the realizability assumption by the bounded rate class noise assumption introduced in the case of passive learning in [6]. More specifically, we assume that there exists a function $f \in F$ such that $\mathbb{P}(Y = f(X)|X) = 1 - \eta(X)$, where $\eta(X) < 1/2$ is the noise rate given X . Since $\eta(X) < 1/2$, the optimal Bayes classifier is in F .

The main technique we use to deal with the noise is applying an adaptive sampler to find out the “true” labels based on the teacher’s noisy answers. In contrast to passive sampling, the sample size in adaptive sampling is a random quantity that may depend on the already seen samples (more technically, a stopping time). Adaptive samplers have been studied before in [10] in more generality, but they give no explicit bounds on the number of samples needed in the special case of interest to us here. To get such, we present a refined and simplified version of their general results that applies to our setting.

Lemma 1. *Suppose we have a coin with an unknown probability p of heads. Let $\delta > 0$. There is an adaptive procedure for tossing the coin such that, with probability at least $1 - \delta$*

1. *The number of coin tosses is at most*

$$\frac{\ln(2/\delta)}{4(1/2 - p)^2} \log \left(\frac{\ln(2/\delta)}{4(1/2 - p)^2} \right) = \tilde{O} \left(\frac{\ln(2/\delta)}{4(1/2 - p)^2} \right).$$

2. *The procedure reports correctly whether heads or tails is more likely.*

```

AdaptiveQuery( $\delta$ )
set  $n_0 = 1$  and toss the coin once;
for  $k = 0, 1, \dots$ 
     $p_k =$  frequency of heads in all tosses so far
     $I_k = [p_k - \sqrt{\frac{(k+1)\ln(2/\delta)}{2^k}}, p_k + \sqrt{\frac{(k+1)\ln(2/\delta)}{2^k}}]$ 
    if( $0.5 \notin I_k$ ) break
    toss the coin  $n_k$  more times, and set  $n_{k+1} = 2n_k$ 
end
if( $I_k \subset [-\infty, 0.5]$ ) output TAILS
else output HEADS

```

Fig. 2. Procedure for determining the more likely outcome of a coin.

Proof. Consider the algorithm of Figure 2. By the Hoeffding bound, $p \in I_k$ with probability at least $1 - \delta/2^{k+1}$, and thus by the union bound the invariant $p \in I_k$ is true for all k with probability at least $1 - \delta$. Provided that this invariant is true, the algorithm clearly cannot output an incorrect answer. And by the same invariant, due to the length of I_k decreasing toward zero, the algorithm will output something after at most the claimed number of coin tosses (in the special case $p = 1/2$ the algorithm will keep tossing the coin indefinitely, but in this case the bound on the number of tosses is also infinite). \square

Note that the adaptive sampler of the above lemma is almost as efficient as passive sampling would be if $|p - 1/2|$ was known in advance. Our positive result presented in the next theorem uses the adaptive sampler as a noise-cancelling subroutine. A similar method for cancelling class noise by repeated queries was independently presented in the query learning context in [7]. However, their strategy uses passive sampling, and thus either requires prior knowledge on $|p - 1/2|$ or a separate step for estimating a lower bound for it. Due to their method needing extra samples in this separate estimation step, our proposed solution will have a smaller total sample complexity.

Theorem 1. *Let A be an active learning algorithm for F that requires $n(\epsilon, \delta)$ label queries and $m(\epsilon, \delta)$ unlabeled examples to be (ϵ, δ) -successful in the realizable case. Then A can be transformed into a noise-tolerant (ϵ, δ) -successful active learner A' for the class of distributions obtained by adding bounded rate class noise to the distributions on which A is successful. With probability at least $1 - \delta$, the unlabeled sample complexity $m'(\epsilon, \delta)$ of A' is $m(\epsilon, \delta/3)$, and if the noise rate is upper bounded by $\alpha < 1/2$, then the label complexity $n'(\epsilon, \delta)$ of A' is at most*

$$n'(\epsilon, \delta) = \tilde{O} \left(\frac{\ln \left(\frac{18\mathbb{E}[n(\epsilon, \delta/3)]}{\delta^2} \right)}{4(1/2 - \alpha)^2} \right) n(\epsilon, \delta/3).$$

Proof. We transform A to A' by replacing each label query $\text{Teacher}(x)$ made by A by a call to **AdaptiveQuery**(δ'), where the role of the coin is played by the teacher that is corrupted by noise. By choosing δ' appropriately, we can ensure that if A does not make too many label queries, then all calls to **AdaptiveQuery** give the correct answer with sufficiently high probability, and thus A' outputs

exactly the same answer in the noisy case as A would have done in the realizable case. We next show how this can be done in detail.

First split δ into three equal parts, covering the three ways in which the modified active learner A' simulating the behavior of A may fail. The simulation A' may fail because A fails in the realizable case, A makes dramatically more label queries than expected, or one or more invocations to the adaptive sampling procedure of Lemma 1 used in the simulation fails. The first case can be covered by setting the parameters of A in the simulation to $(\epsilon, \delta/3)$. For the second case, we use Markov's inequality which implies that the probability of the inequality $n(\epsilon, \delta/3) \leq 3/\delta \cdot \mathbb{E}[n(\epsilon, \delta/3)]$ failing is at most $\delta/3$. In case it does not fail, we have an upper bound for the number of invocations to Lemma 1, and so splitting the remaining $\delta/3$ to the invocations of the adaptive sampler lets us choose its confidence parameter to be $\delta' = \delta^2/(9\mathbb{E}[n(\epsilon, \delta/3)])$. A simple application of the union bound then shows that the total probability of any of the failures happening is at most δ .

In case the no bad event happens, Lemma 1 shows that each of its invocations requires at most

$$\tilde{O}\left(\frac{\ln\left(\frac{18\mathbb{E}[n(\epsilon, \delta/3)]}{\delta^2}\right)}{4(1/2 - \alpha)^2}\right)$$

calls to the noisy teacher. Also, if all these invocations give the correct answer, then A' behaves exactly as A , so the total number of label queries will be the label complexity $n(\epsilon, \delta/3)$ of A in the realizable case times the above, giving the label complexity in the theorem statement. By the same argument of identical behavior, the number of unlabeled examples $m'(\epsilon, \delta)$ required by A' is $m(\epsilon, \delta/3)$.

To complete the proof, it remains to observe that since the noise rate is bounded, the true target $f \in F$ in the realizable case is still the best possible classifier in the bounded class noise rate case. Thus, provided that none of the bad events happens, the fact that A provides an ϵ -approximation to the target in the realizable case directly implies that A' provides an ϵ -approximation to the best function in F in the noisy case. \square

The above theorem shows that allowing bounded rate class noise increases the active learning label complexity only by at most a multiplicative factor determined by the bound on the noise rate α and the logarithm of the label complexity of the active learning algorithm for the realizable case. Thus, for $\alpha < 1/2$ and neglecting logarithmic factors, the order of label complexity as a function of ϵ is unaffected by this kind of noise, so exponentially small label complexity is still possible. As the lower bound presented in the next section shows that the dependence on α is optimal, at most a logarithmic factor could be gained by designing active learners for the bounded rate class noise model directly instead of using the transformation.

Interestingly, it has been recently shown that if the noise rate is not bounded away from $1/2$ but may approach it near the class boundary, then exponential label complexity savings are no longer possible [11]. Thus, relaxing the conditions on the noise in this dimension any more is not possible without sacrificing the

exponential savings: the optimal classifier being in F is not enough, but the noise rate really has to be bounded.

It can be claimed that the way A' deals with class noise is an abuse of the learning and/or noise model, that is, that A' cheats by making repeated queries. It may be, for example, that repeated queries are not possible due to practical reasons (e.g., teacher destroys the objects as a side effect of determining the label). Also, it might be more natural to assume that the teacher makes random errors, but is persistent in the sense that it always gives the same answer when asked the same question. However, such persistently noisy answers define a deterministic labelling rule for all objects, so once the teacher is fixed, there is no randomness left in the noise. Thus, this kind of persistent noise is more naturally dealt with in the model of statistical learning theory that allows true non-realizability.

While the strategy of repeated queries looks suspicious and unlikely to have wide applicability in practice, we believe it is an artifact of suspicious modelling assumptions and should not be prohibited explicitly without additional reasons. It seems to us that even strategies that are not explicitly designed to use repeated queries may actually choose to do so, and thus great care should be taken in their analysis if repetitions are not permissible in the intended applications. As a special case, the number of times an object appears in the pool or stream of unlabeled data should not be automatically taken as an upper bound for the number of queries to the object's label — the original motivation for restricting label queries to unlabeled objects that occur in the sample from P_X was to control the difficulty of the queries [1], and repetitions hardly make a query more difficult. It is also noteworthy that in the regression setting the analogous phenomenon of repeated experiments is more a rule than an exception in optimal solutions to experimental and sequential design problems [12], whereas nonrepeatable experiments are handled as a separate special case [13]. The success of the experimental design approach suggests that maybe there is place for repeatable queries in active learning, too, and that repeatable and nonrepeatable queries definitely deserve separate treatment. While it is unclear to us how the case of nonrepeatable queries can be dealt with efficiently, the next section provides some idea of the difficulties arising there.

4 Negative Result

Let's now move on to true non-realizability and assume only that the learning task P is such that F contains a classifier with a small generalization error of at most β on P . That is, the class of distributions on which we wish the active learner to be successful is

$$\mathcal{P}_{F,\beta} = \{P \mid \exists f \in F : P(f(X) \neq Y) \leq \beta\}.$$

This class allows the target Y to behave completely arbitrarily at least on a set of objects with probability β . A related class of interest to us is

$$\mathcal{P}_{F,\beta}^{\text{det}} = \{P \in \mathcal{P}_{F,\beta} \mid \exists g : P(Y = g(X)) = 1\},$$

in which random noise is excluded by postulating that Y is a function of X . This class models a situation in which the phenomenon to be learned is known to be deterministic, but only approximable by F (e.g., learning the conditions on inputs under which a deterministic computer program crashes).

The role of β is very similar to the role of the bound α on the noise rate in the previous section. While β has a natural interpretation as the best generalization performance achievable by F , it should not be thought of as a parameter known to the learner. Rather, the idea is to study active learning under the unrestricted non-realizability assumption (corresponding to the case $\beta = 1$), and just express the lower bounds of such methods in terms of β .

The question we analyze in this section is the following: Assuming $P \in \mathcal{P}_{F,\beta}$ or $P \in \mathcal{P}_{F,\beta}^{\text{det}}$, what is the expected number of label queries needed to actively learn a classifier from F whose generalization error is with probability at least $1 - \delta$ within ϵ from the optimum in F ?

4.1 Lower Bound for the Noisy Case

In this section we introduce the ideas needed for the lower bound in the case of deterministic non-realizability by considering the simpler case in which random noise is allowed. The special case $\beta = 1/2 - \epsilon$ follows directly from lower bounds for learning with membership queries presented in [9], but the case of general $\beta > 0$ is to our knowledge new even when random noise is allowed.

The problem we study is predicting whether a coin with bias $1/2 \pm \epsilon$ is biased toward heads or tails². This corresponds to the case where $\beta = 1/2 - \epsilon$ and P_X is concentrated on a single point $x_0 \in \mathcal{X}$ on which not all the classifiers in F agree. We further assume that $P(Y|X = x)$ for $x \neq x_0$ is the same for both possibilities of $P(Y|X = x_0)$ and that the learner knows it only has to distinguish between the two remaining alternative distributions P , so queries to objects other than x_0 provide no new information.

Intuitively, it seems clear that an active learner cannot do much here, since there is nothing but x_0 to query and so the only control the learner has is the number of queries. Indeed, by a known result from adaptive sampling mentioned in [10], an active learner still needs an expected number of $\Omega(1/\epsilon^2)$ label queries in this case, and thus has no advantage over passive learning.

The above argument gives a lower bound for the special case $P \in \mathcal{P}_{F,1/2}$, provided $|F| \geq 2$. Adapting the argument for general β can be done as follows. Suppose F contains two classifiers, say, f_0 and f_1 , that sometimes agree and sometimes disagree with each other — this is always true if $|F| > 2$. Place P_X -probability 2β on an object x_0 on which f_0 and f_1 disagree, and the remaining P_X -probability $1 - 2\beta$ on an object x_1 on which they agree. Now, embed the above coin tossing problem with ϵ/β in place of ϵ to the object x_0 on which f_0 and f_1 disagree, and let both f_0 and f_1 be always correct on x_1 . This way, the best classifier in F has error at most β — the better of the classifiers f_0 and f_1

² This learning problem is also a simple example of a case in which prohibiting repeated queries or insisting on persistence of noise makes no sense.

errs at most half the time on x_0 and neither errs on x_1 . By the coin tossing lower bound, $\Omega(\beta^2/\epsilon^2)$ label queries are needed to find out whether f_0 or f_1 is better, even assuming the learner never wastes efforts on querying any other points. As the active learner fails to achieve accuracy ϵ if it chooses incorrectly between f_0 and f_1 , a lower bound $\Omega(\beta^2/\epsilon^2)$ for active learning for $P \in \mathcal{P}_{F,\beta}$ follows.

The above lower bound leaves open the possibility that the difficulties for active learning are caused by high noise rates, not by non-realizability per se. This is a significant weakness, since even though non-realizability can rarely be circumvented in practice, noise-free problems are quite common, e.g., in the verification domain. In such cases, it is reasonable to assume that there really exists a deterministic target, but that it cannot be expected to lie in any sufficiently small F . In the next section, we will extend our lower bound to such cases by essentially derandomizing the arguments outlined above.

4.2 Lower Bound for Deterministic Targets

The lower bound for deterministic targets builds on the techniques used in probing a lower bound for adaptive sampling [14]. In adaptive sampling, the task is to estimate the fraction of inputs that make an unknown boolean function output 1 by querying the values of the function on inputs chosen by an arbitrary adaptive randomized algorithm. Such adaptive samplers can be used, e.g., to estimate the number of rows returned by a query to a database without going through the whole database.

By viewing the unknown boolean function whose bias is to be estimated as the target and assuming P_X is uniform on the domain of this function, it can be seen that adaptive sampling is very close to active learning with queries under uniform distribution. The only difference is that to be (ϵ, δ) -successful, an adaptive sampler is required to output an ϵ -approximation of the bias of the unknown function with probability at least $1 - \delta$, whereas an active learner has to approximate the unknown function and not only its bias.

It is clear that an active learner can be used to solve the seemingly easier task of adaptive sampling, but the other direction that we would need here is less obvious. Hence, we take a different route and instead look directly at the problem that underlies the difficulty of adaptive sampling according to the proof in [14]. Using our terminology, the lower bound for adaptive sampling is proved there through the following result:

Theorem 2. *Let $\epsilon \leq \frac{1}{8}$, $\delta \leq \frac{1}{6}$, and assume the sample complexity n of the sampler is at most $\sqrt{M}/4$ for some large enough M . Let P_X be the uniform distribution on a set \mathcal{X} of size M . Consider target functions*

$$g \in G = \left\{ g: \mathcal{X} \rightarrow \{0, 1\} \mid P_X(g(X) = 1) = \frac{1}{2} \pm \epsilon \right\}$$

Then any randomized adaptive sampling algorithm for the task of determining the bias of any such target $g \in G$ correctly with probability at least $1 - \delta$ requires on some $g \in G$ an expected sample complexity of at least $\Omega\left(\frac{1}{\epsilon^2} \ln \frac{1}{\delta}\right)$.

Theorem 2 implies a lower bound for adaptive sampling, since an (ϵ, δ) -successful adaptive sampler solves the above decision problem as a special case. However, in case of active learning, we need a strengthened version of Theorem 2 that is fortunately also implied by exactly the same proof presented in [14] (see especially the beginning of the proof of Theorem 1 therein). Namely, the proof in [14] shows that the lower bound applies also to the version of the decision problem in which the adaptive sampler is required only to give the correct answer with probability at least $1 - \delta$ with respect to its internal randomness *and* the choice of g from the uniform distribution on G . To solve this modified problem, even the simplistic strategy of first running an active learner that outputs an f and then predicting that $P_X(g(X) = 1) = 1/2 + \epsilon$ iff $P_X(f(X) = 1) > 1/2$ suffices. More exactly, it can be shown that the probability of the event that the bias of the learned f differs from that of the randomly chosen target $g \in G$ is at most δ . This is accomplished by the following lemma, in which the boundedness assumption on label complexity can be replaced by an application of Markov's inequality if desired.

Lemma 2. *Let A be an $(\epsilon/2, \delta/2)$ -successful active learning algorithm for $\mathcal{P}_{F, 1/2-\epsilon}^{det}$, where F contains the constant classifiers 0 and 1. Suppose that the label complexity $n(\epsilon/2, \delta/2)$ of A is bounded by $N < \infty$. Suppose \mathcal{X} is infinite and P_X is a uniform distribution on a sufficiently large finite set $\mathcal{X}_0 \subset \mathcal{X}$. Then, the probability that the bias of the classifier f output by A is the same as the bias of the target chosen uniformly at random from*

$$G_0 = \{g: \mathcal{X} \rightarrow \{0, 1\} \mid P_X(g(X) = 1) = \frac{1}{2} \pm \epsilon \text{ and } g(x) = 0 \text{ for } x \notin \mathcal{X}_0\}$$

is at most δ .

Proof (Sketch). As F contains the constant classifiers 0 and 1, the minimal generalization error achievable using F is always at most $1/2 - \epsilon$ on any target $g \in G_0$. Thus, for each such $g \in G_0$, A can output a classifier f with generalization error larger than $1/2 - \epsilon/2$ on the learning problem defined by P_X and g with probability at most δ . Since this holds for each $g \in G_0$, we can let g be chosen randomly to get $\mathbb{P}(\epsilon(f) \geq 1/2 - \epsilon/2) \leq \delta$, where \mathbb{P} now denotes probability with respect to randomness in A (internal and that caused by the random choice of U if such is used) and the choice of g from the uniform distribution over G_0 .

Let B denote whether a classifier is biased toward 0 or 1, i.e., $B(f) = 1$ iff $P_X(f(X) = 1) \geq 1/2$ and $B(f) = 0$ otherwise. Conditioning on whether $B(f) = B(g)$ or not, we get from the previous inequality that

$$\begin{aligned} \delta &\geq \mathbb{P}\left(\epsilon(f) \geq \frac{1}{2} - \frac{\epsilon}{2}\right) = \mathbb{P}\left(\epsilon(f) \geq \frac{1}{2} - \frac{\epsilon}{2} \mid B(f) = B(g)\right) \mathbb{P}(B(f) = B(g)) \\ &\quad + \mathbb{P}\left(\epsilon(f) \geq \frac{1}{2} - \frac{\epsilon}{2} \mid B(f) \neq B(g)\right) \mathbb{P}(B(f) \neq B(g)) \\ &\geq \mathbb{P}\left(\epsilon(f) \geq \frac{1}{2} - \frac{\epsilon}{2} \mid B(f) \neq B(g)\right) \mathbb{P}(B(f) \neq B(g)), \end{aligned}$$

implying $\mathbb{P}(B(f) \neq B(g)) \leq \frac{\delta}{\mathbb{P}(\varepsilon(f) \geq 1/2 - \epsilon/2 | B(f) \neq B(g))}$. Thus, to show that the probability of the event $B(f) \neq B(g)$ — that is, the event that using the strategy of predicting that the bias of g is that of f fails — has probability at most 2δ , it suffices to show that

$$\mathbb{P}(\varepsilon(f) \geq 1/2 - \epsilon/2 | B(f) \neq B(g)) \geq 1/2. \quad (1)$$

We outline an argument showing this next.

Let $S \subset \mathcal{X}$ denote the (random) set of points that A queries. We can assume that $S \subset \mathcal{X}_0$, since points outside \mathcal{X}_0 provide no information about g and can thus be ignored in the analysis. Also, since we assume $|S| \leq N$, we can take \mathcal{X}_0 so large that fraction of $|S|/|\mathcal{X}_0|$ is arbitrarily small. Note that f can depend on the values of the target g on S only, but not on its values outside S as these are never observed by A . Thus, for each choice of f , S , and the values of g on S observed by A , the number of errors f makes on $\mathcal{X}_0 \setminus S$ is a random variable whose distribution is induced by the distribution of g conditioned on the values of g on S and the underlying conditioning event $B(g) \neq B(f)$.

Now we apply the fact that \mathcal{X}_0 can be made so large in comparison to S that the values of g and f on S do not affect their biases on $\mathcal{X}_0 \setminus S$ by much. Consider any $x \in \mathcal{X}_0 \setminus S$. If $f(x) = B(f)$, the probability over the choice of g of the event $f(x) \neq g(x)$ is (about) $1/2 + \epsilon$, and if $f(x) \neq B(f)$, the corresponding probability is (about) $1/2 - \epsilon$. Thus, since the former case is more probable provided \mathcal{X}_0 is large, the expected error of f on $\mathcal{X}_0 \setminus S$ is at least almost $1/2$. The contribution of the error of f on S to its overall error can be made arbitrarily small again by taking \mathcal{X}_0 large enough, from which it follows that the expected generalization error of f is, say, at least $1/2 - \epsilon/4$ for large enough \mathcal{X}_0 . Furthermore, by noting that the distribution of the error of f on $\mathcal{X}_0 \setminus S$ can be expressed in terms of a hypergeometric distribution whose variance goes to zero as \mathcal{X}_0 is increased, it finally follows by Chebysev's inequality that the probability of the error of f being larger than $1/2 - \epsilon/2$ can be made to be at least $1/2$. The details in all these arguments can be made precise by filling in the calculations on how much the behavior of f and g on S can affect their behavior outside S given that $|S|/|\mathcal{X}_0|$ is small, but we omit the tedious details in this draft.

Since the above is true whatever set S the algorithm A decides to query, whatever answers it receives, and what classifier f it chooses, we get inequality (1), finally concluding the proof. \square

The above lemma shows how an $(\epsilon/2, \delta/2)$ -successful active learner for the class of distributions $\mathcal{P}_{F, 1/2-\epsilon}^{\text{det}}$ can be used to solve the average-case version of the decision problem of Theorem 2 discussed after the theorem statement, provided that F contains the constant classifiers 0 and 1. This assumption can be replaced by assuming F contains any two classifiers that are complements of each other, since detecting which of these is closer to the target is equivalent to detecting the bias. Furthermore, we can move the target to within β of F by the same trick we used in Section 3 by embedding the bias detection problem to a subset of \mathcal{X} that has probability 2β , and putting the rest of the probability mass on

a point on which the classifiers agree. These steps together give us the desired lower bound stated below:

Theorem 3. *Let A be an active learning algorithm. Let $\epsilon \leq 1/8$, $\delta \leq 1/6$, suppose the label complexity $n(\epsilon, \delta)$ of A is uniformly bounded for each (ϵ, δ) , and let the unlabeled label complexity $m(\epsilon, \delta)$ be arbitrary. Suppose F is such that it contains two functions f_0, f_1 that agree at least on one point and disagree on infinitely many points. Let $\beta > 0$ and suppose A is successful for $\mathcal{P}_{F, \beta}^{\text{det}}$. Then, for some P_X and target function g , the expected number of $n(\epsilon, \delta)$ is*

$$\Omega\left(\frac{\beta^2}{\epsilon^2} \ln \frac{1}{\delta}\right).$$

The lower bound applies to all active learning algorithms and is not specific to, say, empirical risk minimization, or to any specific hypothesis class. Also, it is easy to see by replacing the point masses in \mathcal{X}_0 by a partition of $\{x \in \mathcal{X} \mid f_0(x) \neq f_1(x)\}$ into equiprobable sets that the lower bound immediately extends to a wide range of distributions P_X (including, e.g., all continuous distributions), and remains true even if P_X is known in advance. Thus, the lower bound does not result from any specific properties of P_X , and cannot be circumvented by any amount of unlabeled data. As an interesting special case, the lower bound applies to learning linear separators with uniform distribution, the best known case in which exponential savings are possible under the realizability assumption. However, the lower bound becomes interesting only when $\epsilon \ll \beta$, so it does not rule out exponential speed-ups in the low accuracy regime. This fits perfectly together with the label complexity upper bounds for an active learning algorithm for linear thresholds presented in [4], where it is shown that exponential speed-ups are indeed possible when $\epsilon > \text{const} \cdot \beta$, after which their upper bound on the learning speed degrades to match the above lower bound (up to constants).

The fact that the lower bound depends on β is unavoidable, since if the target is only slightly outside F , the learner will with high probability fail to even notice the non-realizability. This case is of real importance when using the covering approach for active learning in the realizable case [3]. Making the covering finer as a function of ϵ corresponds to enlarging the covering of the underlying F that the algorithm uses as its hypothesis class so that β goes to zero as the accuracy requirements get stricter. This eliminates the effects of the lower bound, but is of course only possible if we know a suitably small class F for which the problem is realizable in advance. In case the target is truly unknown, circumventing the lower bound by extending F is not possible.

5 Conclusions

We have shown that bounded rate class noise can be relatively easily dealt with by using repeated label queries to cancel the effects of the noise, but that in the truly non-realizable case active learning does not give better rates of sample complexity than passive learning when only the dependence on the accuracy

and confidence parameters is considered. However, even though the lower bound rules out exponential savings in the non-realizable case, the bound leaves open the possibility of reducing the label complexity by at least a factor of β^2 or more as the complexity of F is not reflected in the lower bound. In practice, even such savings would be of great value. Thus, the lower bound should not be interpreted to mean that active learning does not help in reducing the label complexity in the non-realizable case. Instead, the lower bound only means that the reductions will not be exponential, and that the goal of active learning should be readjusted accordingly.

The results in this paper are only a first step toward a full understanding of the label complexity of active learning under various noise models. In particular, it would be interesting to see how the complexity of F and other kinds of noise (noise in objects, malicious noise, . . .) affect the active learning label complexity.

References

1. Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.
2. Sanjoy Dasgupta, Adam Tauman Kalai, and Claire Monteleoni. Analysis of perceptron-based active learning. In *COLT'05*, pages 249–263. Springer-Verlag, 2005.
3. Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS'05*, 2005.
4. Nina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *ICML*, 2006. Accepted.
5. Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2002.
6. Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1987.
7. Yasubumi Sakakibara. On learning from queries and counterexamples in the presence of noise. *Information Processing Letters*, 37(5):279–284, 1991.
8. Vladimir N. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, 1982.
9. Claudio Gentile and David P. Helmbold. Improved lower bounds for learning from noisy examples: an information-theoretic approach. In *COLT'98*, pages 104–115. ACM Press, 1998.
10. Carlos Domingo, Ricard Gavaldà, and Osamu Watanabe. Adaptive sampling methods for scaling up knowledge discovery algorithms. In *DS'99*, pages 172–183. Springer-Verlag, 1999.
11. Rui Castro, March 2006. Personal communication.
12. Samuel D. Silvey. *Optimal Design*. Chapman and Hall, London, 1980.
13. Gustaf Elfving. Selection of nonrepeatable observations for estimation. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 69–75, 1956.
14. Ran Canetti, Guy Even, and Oded Goldreich. Lower bounds for sampling algorithms for estimating the average. *Information Processing Letters*, 53(1):17–25, 1995.