, , 1–38 ()

Selective sampling using the Query by Committee algorithm

YOAV FREUND yoav@research.att.com
AT&T Labs, Murray Hill, NJ 07974

H. SEBASTIAN SEUNG seung@bell-labs.com
Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974

ELI SHAMIR AND NAFTALI TISHBY {shamir,tishby}@cs.huji.ac.il
Institute of Computer Science, Hebrew University, Jerusalem, ISRAEL

Received May 1, 1991

Editor: David Haussler

Abstract. We analyze the "query by committee" algorithm, a method for filtering informative queries from a random stream of inputs. We show that if the two-member committee algorithm achieves information gain with positive lower bound, then the prediction error decreases exponentially with the number of queries. We show that, in particular, this exponential decrease holds for query learning of perceptrons.

Keywords: selective sampling, query learning, Bayesian Learning, experimental design

1. Introduction

Most of the research on the theory of learning from random examples is based on a paradigm in which the learner is both trained and tested on examples drawn at random from the same distribution. In this paradigm the learner is passive and has no control over the information that it receives. In contrast, in the *query* paradigm, the learner is given the power to ask questions. What does the learner gain from this additional power?

Study of the use of queries in learning (Valiant,1984, Angluin,1988), has mostly concentrated on algorithms for *exact identification* of the target concept. This type of analysis concentrates on the worst case behavior of the algorithm, and no probabilistic assumptions are made. In contrast, we are interested in algorithms that achieve approximate identification of the target, and our analysis is based on probabilistic assumptions. We assume that both the examples and the target concept are chosen randomly. In particular, we show that queries can help *accelerate* learning of concept classes that are already learnable from just unlabeled data.

This question was previously studied by (Eisenberg & Rivest,1990) in the PAC learning framework. They give a negative result, and show that, for a natural set of concept classes, which they call "dense in themselves", queries are essentially useless. They show that giving the learner the ability to ask membership queries (questions of the type "what is the label of the point x?") in this context does not enable the learner to significantly reduce the total number of labeled examples it needs to observe. The reason is that if the learner observes only a small number of examples, *either passively or actively*, then it can not be sensitive to slight changes in the target concept and in the underlying distribution. An adversary can alter the distribution and the target in a way that will not cause the learner to change its hypothesis, but will increase the error of this hypothesis in a significant way. In this paper we show how some concept classes that are dense in themselves can be learned efficiently if we allow the learner access to random *unlabeled* examples. This added capability enables the learner to maintain its sensitivity to the input distribution, while reducing the number of labels that it needs to know.

Baum(Baum,1991), proposed a learning algorithm that uses membership queries to avoid the intractability of learning neural networks with hidden units. His algorithm is proved to work for networks with at most four hidden units, and there is experimental evidence(Baum & Lang,1992) that it works for larger networks. However, when Baum and Lang tried to use this algorithm to train a network for classifying handwritten characters, they encountered an unexpected problem (Baum & Lang,1992). The problem was that many of the images generated by the algorithm as queries did not contain any recognizable character, they were artificial combinations of character images that had no natural meaning. The learning algorithm that is analyzed in this paper uses random unlabeled instances as queries and in this way may avoid the problem encountered by Baum's algorithm.

In the lines of work described above, queries are explicitly constructed. In contrast, our work is derived within the *query filtering* paradigm. In this paradigm, proposed by (Cohn, Atlas & Ladner,1990), the learner is given access to a stream of inputs drawn at random from the input distribution. The learner sees every input, but chooses whether or not to query the teacher for the label. Giving the learner easy access to unlabeled random examples is a very reasonable assumption in many real-life contexts. In applications such as speech recognition, it is often the case that collecting unlabeled data is a highly automatic process, while finding the correct labeling of the data requires expensive human work. Our algorithm uses all of the unlabeled examples and in this way overcomes the problems pointed out by Rivest and Eisenberg. Learning becomes an interactive process: rather than requesting the human to label all the examples in advance, we let the computer choose the examples whose labels are most informative. Initially, most examples will be informative for the learner, but as the process continues, the prediction capabilities of the learner improve, and it discards most of the examples as non-informative, thus saving the human teacher a large amount of work.

In (Cohn, Atlas & Ladner,1990) there are several suggestions for query filters together with some empirical tests of their performance on simple problems. In (Seung, Opper & Sompolinsky,1992) the authors suggested a filter called "query by committee," (**QBC**) and analytically calculated its performance for some perceptron-type learning problems. For these problems, they found that the prediction error decreases exponentially fast in the number of queries. In this work we present a more complete and general analysis of query by committee, and show that such an exponential decrease is guaranteed for a general class of learning problems.

The problem of selecting the optimal examples for learning is closely related to the problem of experimental design in statistics (see e.g. (Fedorov,1972, Atkinson,1992)). Experimental design is the analysis of methods for selecting sets of experiments, which correspond to membership queries in the context of learning theory. The goal of a good design is to select experiments in a way that their outcomes, which correspond to labels, give sufficient information for constructing a hypothesis that maximizes some criterion of accuracy. One natural criterion is the accuracy with which the parameters that define the hypothesis can be estimated (Lindley,1956). In the context of Bayesian estimation a very general measure of the quality of a query is the reduction in the entropy of the posterior distribution that is induced by the answer to the query. Similar suggestions have been made in the perceptron learning literature(Kinzel & Ruján,1990). A different experimental design criterion is the accuracy with which the outcome of future experiments, chosen from some constrained domain, can be predicted using the hypothesis. This criterion is very similar to criteria used in learning theory. Both criteria are important for us in this paper. We show that while in the general case the two are not necessarily related, they are related in the case of the query by committee algorithm. Using this relation we prove the efficiency of the algorithm for some specific concept classes.

The results presented in this paper are restricted to a rather limited set of learning problems. The main restriction is that the concepts are assumed to be deterministic and noiseless. In the summary we list what we think are the natural extensions of our analysis.

The paper is organized as follows. In Section 2 we present the Bayesian framework of learning within which we analyze our algorithm. In Section 3 we present some simple learning problems and demonstrate a case in which the information gain of a query is not the relevant criterion when we are interested in prediction quality. In Section 4 we describe the query by committee algorithm. In Section 5 we prove that there is a close relation between information gain and prediction error for **QBC**. Using this relation we show in Section 6 that the prediction error decreases exponentially fast with the number of queries for some natural learning problems. In Section 7 we give a broader view on using unlabeled examples for accelerating learning, and in Section 8 we summarize and point to some potential future directions.

2. Preliminaries

We work in a Bayesian model of concept learning (Haussler, Kearns & Schapire,1994). As in the PAC model, we denote by X an arbitrary sample space over which a distribution $\mathcal{D}$ is defined. In this paper we concentrate on the case where X is a Euclidean space R^d. Each concept is a mapping $c : X \rightarrow \{0,1\}$ and a concept class $\mathbf{C}$ is a set of concepts. The Bayesian model differs from the PAC model in that we assume that the target concept is chosen according to a *prior distribution* $\mathcal{P}$ over

$\mathbf{C}$ and that this distribution is known to the learner. We shall use the notation $\Pr_{x\in\mathcal{D}}(\cdot)$ to denote the probability of an event when x is chosen at random from X according to $\mathcal{D}$.

We assume that the learning algorithm has access to two oracles: **Sample** and **Label**. A call to **Sample** returns an unlabeled example $x \in X$, chosen according to the (unknown) distribution $\mathcal{D}$. A call to **Label** with input x, returns $c(x)$, the label of x according to the target concept. After making some calls to the two oracles, the learning algorithm is required to output a hypothesis $h : X \rightarrow \{0,1\}$. We define the expected error of the learning algorithm as the probability that $h(x) \neq c(x)$, where the probability is taken with respect to the distribution $\mathcal{D}$ over the choice of x, the distribution $\mathcal{P}$ over the choice of c and any random choices made as part of the learning algorithm or of the calculation of the hypothesis h. We shall usually denote the number of calls that the algorithm makes to **Sample** by m and the number of calls to **Label** by n. Our goal is to give algorithms that achieve accuracy ϵ after making $O(1/\epsilon)$ calls to **Sample** and $O(\log 1/\epsilon)$ calls to **Label**.

In our analysis we find it most convenient to view the finite number of instances that are observed by the learning algorithm as an initial segment of an infinite sequence of examples, all drawn independently at random according to $\mathcal{D}$. We shall denote such a sequence of unlabeled examples by $\vec{X} = \{x_1, x_2 \ldots\}$, and use $\langle \vec{X}, c(\vec{X})\rangle = \{\langle x_1, c(x_1)\rangle, \langle x_2, c(x_2)\rangle \ldots\}$ to denote the sequence of labeled examples that is generated by applying c to each $x \in \vec{X}$. We use $\vec{X}_{1\ldots m}$ to denote the sequence of the first m elements in $\vec{X}$. We use the terminology of (Mitchell,1982), and define the *version space* generated by the sequence of labeled examples $\langle \vec{X}_{1\ldots m}, c(\vec{X}_{1\ldots m})\rangle$ to be the set of concepts $c' \in \mathbf{C}$ that are consistent with c on $\vec{X}$, i.e. that $c'(x_i) = c(x_i)$ for all $1 \leq i \leq m$. We denote the version space that corresponds to the first i labeled examples by $V_i = V(\langle \vec{X}_{1\ldots i}, c(\vec{X}_{1\ldots i})\rangle)$. The initial version space, $V_0 = V(\emptyset)$, is equal to $\mathbf{C}$. The version space is a representation of the information contained in the set of labeled examples observed by the learning algorithm. A natural measure of the progress of the learning process is the rate at which the size of the version space decreases. The *instantaneous information gain* from the ith labeled example in a particular sequence of examples is defined to be $-\log \Pr_{\mathcal{P}}(V_i)/\Pr_{\mathcal{P}}(V_{i-1})$. Summing the instantaneous information gains over a complete sequence of examples we get the *cumulative information gain*, which is defined as

$$\mathcal{I}(\langle x_1, c(x_1)\rangle, \ldots, \langle x_m, c(x_m)\rangle) \doteq -\sum_{i=1}^{m} \log \frac{\Pr_{\mathcal{P}}(V_i)}{\Pr_{\mathcal{P}}(V_{i-1})} = -\log \Pr_{\mathcal{P}}(V_m) \ . \quad (1)$$

The natural measure of the information that we expect to gain from the label of an unlabeled example is the expected instantaneous information gain taken with respect to the probability that each one of the two labels occurs. Let p_0 be the probability that the label of x_m is 0, given that $c \in V_{m-1}$ and let V_m^0 be the version space that results from the label x_m being 0. Define p_1 and V_m^1 in the corresponding way for the case $c(x_m) = 1$. We define the *expected information gain* of x_i, given

V_{i-1}, to be:

$$\begin{aligned}\mathcal{G}(x_i|V_{i-1}) &\doteq -p_0 \log \frac{\Pr_{\mathcal{P}}(V_i^0)}{\Pr_{\mathcal{P}}(V_{i-1})} - p_1 \log \frac{\Pr_{\mathcal{P}}(V_i^1)}{\Pr_{\mathcal{P}}(V_{i-1})} \\ &= -p_0 \log p_0 - (1-p_0)\log(1-p_0) \doteq \mathcal{H}(p_0) \ ,\end{aligned} \tag{2}$$

where $\mathcal{H}(p)$ denotes the Shannon information content of a binary random variable whose probability of being 1 is p. We shall use log base 2 in our definition and measure the expected information gain in *bits*.[1] The maximal information gain from a single label is one bit. The information gain is thus a very attractive measure of the gain that can be expected from asking **Label** for the label of an example. However, as we show in Section 3, this measure, by itself, is not sufficient for guaranteeing a large reduction in the expected prediction error of the algorithm.

The "Gibbs" prediction rule is to predict the label of a new example x by picking a hypothesis h at random from the version space and labeling x according to it. The random choice of h is made according to the prior distribution $\mathcal{P}$ restricted to the version space. It is a simple observation (see (Haussler, Kearns & Schapire,1994)), that the expected error of this prediction error is at most twice larger than the expected error of the optimal prediction rule which is the Bayes rule. We shall assume that our learning algorithm has access to an oracle, denoted **Gibbs**, which can compute the Gibbs prediction for a given example $x \in X$ and version space $V \subset \mathbf{C}$. Each time $\mathbf{Gibbs}(V, x)$ is called, a hypothesis $h \in \mathbf{C}$ is chosen at random according to the distribution $\mathcal{P}$ restricted to V, and the label $h(x)$ is returned. Note that two calls to **Gibbs** with the same V and x can result in different predictions. The main result of the paper is that a simple algorithm for learning using queries, that uses the Gibbs prediction rule, can learn some important concept classes with accuracy that is exponentially small in the number of calls to **Label**.

3. Two simple learning problems

In this section we discuss two very simple learning problems. Our goal here is to give examples of the concepts defined in the previous section and to show that constructing queries solely according to their expected instantaneous information gain is not a good method in general.

Consider the following concept class. Let $X = [0,1]$, and let the associated probability distribution $\mathcal{D}$ be the uniform distribution. Let the concept class $\mathbf{C}$, consist of all functions of the form

$$c_w(x) = \begin{cases} 1, & w \le x \\ 0, & w > x \end{cases} \ , \tag{3}$$

where $w \in [0,1]$. We define the prior distribution of concepts, $\mathcal{P}$ to be the one generated by choosing w uniformly from $[0,1]$.

The version space defined by the examples $\{\langle x_1, c(x_1)\rangle, \ldots, \langle x_m, c(x_m)\rangle\}$ is (isomorphic to) the segment $V_i = [\max(x_i|c(x_i)=0), \min(x_i|c(x_i)=1)]$. Let us denote

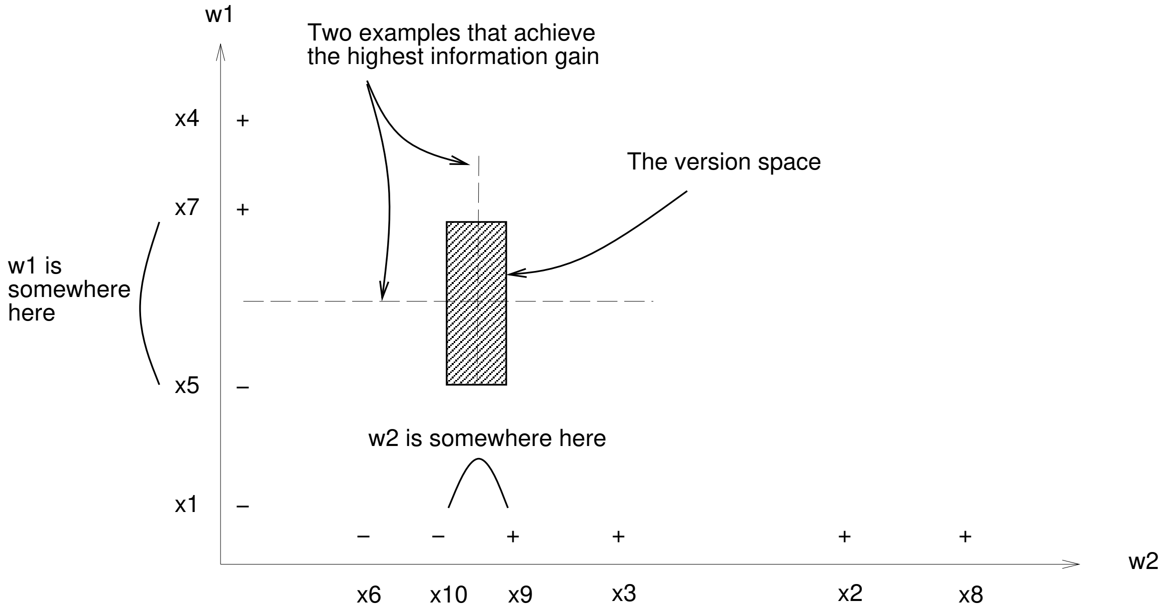

Figure 1. A figure of the version space and the examples that achieve maximal information gain for the two threshold learning problem defined below.

by ξ_i the ratio of the probabilities of the version space before and after observing the ith example, i.e. $\xi_i = \Pr_{\mathcal{P}} V_i / \Pr_{\mathcal{P}} V_{i-1}$. The instantaneous information gain of the example $\langle x_i, c(x_i) \rangle$ is $-\log \xi_i$. Given an *unlabeled* example, the expected instantaneous information gain from x_i is $\mathcal{H}(\xi_i)$. Examples that fall outside the segment have zero expected information gain, while the example that divides the segment into two equal parts obtains the highest possible expected information gain of one bit. This agrees with our intuition because the labels of examples that fall outside the segment are already completely determined by previous labeled examples, while the label of the example that falls in the middle of the version space interval is least predictable. It is easy to show that the probability of a prediction error for the Gibbs prediction rule is equal to the length of the segment divided by three. Thus, if the learner asks for the label of the example located in the middle of the segment, it is guaranteed to halve the error of the Gibbs prediction rule. In this case we see that asking the oracle **Label** to label the example that maximizes the expected information gain guarantees an exponentially fast decrease in the error of the Gibbs prediction rule. In contrast, the expected prediction error after asking for the labels of n randomly chosen examples is $O(1/n)$.

The question is whether constructing queries according to their expected information gain is a good method in general, i.e. whether it always guarantees that the prediction error decreases exponentially fast to zero.

The answer to this question is negative, to see why this is the case consider the following, slightly more complex, learning problem. Let the sample space be the set of pairs in which the first element, i, is either 1 or 2, and the second element, z, is a real number in the range $[0,1]$, i.e. $x \in X = \{1,2\} \times [0,1]$. Let $\mathcal{D}$ be the distribution defined by picking both i and z independently and uniformly at random. Let the concept class be the set of functions of the form

$$c_{\vec{w}}(i,z) = \begin{cases} 1, & w_i \leq z \\ 0, & w_i > z \end{cases}, \tag{4}$$

where $\vec{w} \in [0,1]^2$. The prior distribution over the concepts is the one generated by choosing $\vec{w}$ uniformly at random from $[0,1]^2$. In this case each example corresponds to either a horizontal or a vertical half plane, and the version space, at each stage of learning, is a rectangle (see Figure 3). There are always two examples that achieve maximal information gain, one horizontal and the other vertical. Labeling each one of those examples reduces the volume of the version space by a factor of two. However, the probability that the Gibbs rule makes an incorrect prediction is proportional to the perimeter of the rectangular version space, and not to its volume. Thus, if the learner always constructs queries of the same type, only one of the dimensions of the rectangle is reduced, and the perimeter length stays larger than a constant. This implies that the prediction error also stays larger than a constant.

We conclude that the expected information gain of an unlabeled example is *not* a sufficient criterion for constructing good queries. The essential problem is that the distribution over the examples is completely ignored by this criterion. While one can easily find a specific solution for the given learning problem, we would like to have a general method that is sensitive to the distribution of the examples, and is guaranteed to work for a wide variety of problems. In the next section we present such a method.

4. The Query by Committee learning algorithm

In (Seung, Opper & Sompolinsky,1992) the authors devise an algorithm for learning with queries which they called "Query by Committee" and we shall refer to as the **QBC** algorithm. The algorithm uses as queries examples whose expected information gain is high, however, rather than *constructing* the examples, it *filters* the more informative examples from the random unlabeled examples that it gets from the oracle **Sample**. We discuss the simplest case in which the committee is of size two. [2]

The algorithm proceeds in iterations. In each iteration it calls **Sample** to get a random instance x. It then calls **Gibbs** twice, and compares the two predictions for the label of x. If the two predictions are equal, it rejects the instance and proceeds to the next iteration. If the two predictions differ, it calls **Label** with input x, and adds the labeled example to the set of labeled examples that define the version space.

It then proceeds to the next iteration. In (Seung, Opper & Sompolinsky,1992) the authors treat the query by committee algorithm as an on-line learning algorithm, and analyze the rate at which the error of the two Gibbs learners reduces as a function of the number of queries made. In our work we prove general bounds both on the number of queries and on the number of random examples that the algorithm tests. In order to do that we consider a *batch* learning scenario, in which the learning algorithm is tested only after it has finished observing all of the training examples and has fixed its prediction hypothesis.

To do that we define a termination condition on the iterative process described above. When the algorithm reaches this a state that fulfills this condition it stops calling **Sample** and **Label** and uses the **Gibbs** oracle to *predict* the labels of the instances that it receives in the test phase. The termination condition is satisfies if a large number of consecutive instances supplied by **Sample** are all rejected.

We measure the quality of the predictions made by the algorithm in a way similar to that used in Valiant's PAC model. We define the expected error of the algorithm as the probability that its prediction of the label of a random instance disagrees with that of the true underlying concept. This probability is taken with respect to the random choice of the instance as well as the underlying concept. We also allow the algorithm some small probability of failure to account for the fact that the sequence of instances that it observes during training is atypical.

We say that the learning algorithm is successful if its expected error is small, when trained on a typical sequence of instances. More precisely, we define two parameters, an accuracy parameter $1 > \epsilon > 0$ and a confidence parameter $1 > \delta > 0$. We use the term "training history" to describe a specific sequence of random instances and random coin flips used during learning a specific hidden concept. For each choice of the hidden concept, we allow a set of training histories that has probability δ to be marked as "atypical" training histories. Our requirement is that the expected error over the set of typical training histories is smaller than ϵ. The parameters ϵ and δ are provided to the learning algorithm as input and are used to define the termination criterion. Figure 2 gives a formal description of the algorithm. It is important to notice that the termination condition depends only on ϵ and δ, and not of any properties of the concept class. While the performance of the algorithm *does depend* on such properties, the algorithm can be used without prior knowledge of these properties.

It is easy to show that if **QBC** ever stops, then the error of the resulting hypothesis is small with high probability. That is because it is very unlikely that the algorithm stops if the probability of error is larger than ϵ (proof is given in Lemma 2). The harder question is whether **QBC** ever stops, and if it does, how many calls to **Sample** and to **Label** does it make before stopping? As we shall show in the following two sections, there is a large class of learning problems for which the algorithm will stop, with high probability, after $O(1/\epsilon \ \log 1/\delta\epsilon)$ calls to **Sample**, and $O(\log 1/\epsilon)$ calls to **Label**.

The committee filter tends to select examples that split the version space into two parts of comparable size, because if one of the parts contains most of the version

Input: $\epsilon > 0$ - the maximal tolerable prediction error.
$\delta > 0$ - the desired reliability.
Gibbs- an oracle that computes Gibbs predictions.
Sample- an oracle that generates unlabeled examples.
Label- an oracle that generates the correct label of an example.

Initialize n - the counter of calls to **Label** – to 0, and set the initial version space, V_0, to be the complete concept class **C**.

Repeat until more than t_n consecutive examples are rejected. Where

$$t_n = \frac{1}{\epsilon} \ln \frac{\pi^2(n+1)^2}{3\delta} ,$$

and n is the number of examples that have been used as queries so far.

1. Call **Sample** to get an unlabeled example $x \in X$ drawn at random according to $\mathcal{D}$.
2. Call **Gibbs**(V_n, x) twice, to get two predictions for the label of x.
3. **If** the two predictions are equal **then** reject the example and return to the beginning of the loop. (step 1)
4. **Else** call **Label**(x) to get $c(x)$, increase n by 1, and set V_n to be all concepts $c' \in V_{n-1}$ such that $c'(x) = c(x)$.

Output as the prediction hypothesis **Gibbs**(V_n, x).

Figure 2. Query by a committee of two

space, then the probability that the two hypotheses will disagree is very small. Let us normalize the probability of the version space to one and assume that an example x partitions the version space into two parts with probabilities F and $1 - F$ respectively. Then the probability of accepting the example x as a query is $2F(1 - F)$ and the information gain from an example is $\mathcal{H}(F)$. . Both of these functions are maximized at $F = 0.5$ and decrease symmetrically to zero when F is increased to one or decreased to zero. It is thus clear that the queries of **QBC** have a higher expected information gain than random examples. However, it is not true in general that the expected information gain of the queries will always be larger than a constant,[3] moreover, as we have seen in the Section 3, queries with high information gain do not guarantee a fast decrease of the prediction error in general. Our proof of the performance of **QBC** consists of two parts. In the first part, given in Section 5, we show that a lower bound on the information gain of the queries *does* guarantee a fast decrease in the prediction error of **QBC**. In the second part, given in Section 6, we show that the expected information gain of the queries of **QBC** is guaranteed to be higher than a constant in some important cases.

5. Relating information gain and prediction error for Query by Committee

In this section we prove that if the expected information gain from the queries used by **QBC** is high, then the prediction error of the algorithm is guaranteed to be exponentially small in the number of queries asked. We shall first define exactly what we mean by high information gain, and then give the theorem and its proof.

In our analysis we treat runs of the algorithm as initial segments of infinite runs that would have been generated had there been no termination criterion on the execution of the main loop in **QBC**. We denote by $\vec{X}$ the infinite sequence of unlabeled examples that would have been generated by calls to **Sample**. We use an infinite sequence of integer numbers $I = \{1 \leq i_1 < i_2 < \ldots\}$ to refer to the sequence of indices of those examples that are filtered by **QBC** from $\vec{X}$ and used as queries to **Label**. This set of examples is denoted $\vec{X}_I$. We denote by M the sequence of integers from 1 to m, and use $\vec{X}_M$ to denote the first m examples in $\vec{X}$. We use I_n to denote the first n elements of I. Finally, $\vec{X}_{I_n}$ indicates the first n examples that are used as queries, and $\vec{X}_{I\cap M}$ indicates the queries that are chosen from the first m unlabeled examples.

We now present the probabilistic structure underlying the query process. A point in the sample space Ω is a triple $\langle c, \vec{X}, I\rangle$. The probability distribution over this space is defined as follows. The target concept c is chosen according to $\mathcal{P}$, and each component in the infinite sequence $\vec{X}$ is chosen independently according to $\mathcal{D}$. Fixing c and $\vec{X}$, we define the distribution of the first n elements of I according to the probability that algorithm **QBC** calls the oracle **Label** on the iterations indexed by I_n. It is easy to see that the distributions defined for different values of n are consistent with each other, thus we can define the distribution on I as the limiting distribution for $n \to \infty$. We denote the distribution we have defined on the triplets $\langle c, \vec{X}, I\rangle$ by Δ and use $\Pr_\Delta$ and E_Δ to indicate the probability and the expectation taken with respect to this distribution.

We now define formally what we mean when we say that the queries of **QBC** are informative.

Definition. We say that the expected information gain of queries made by **QBC** for the learning problem of concept class **C**,concept distribution $\mathcal{P}$, and input distribution $\mathcal{D}$, is uniformly lower bounded by $g > 0$ if the following holds.

For the distribution over $\langle c, \vec{X}, I\rangle$ that is generated by $\mathcal{P}, \mathcal{D}$ and **QBC** and for every $n \geq 0$, the expected instantaneous information gain from the $n+1$st query, given any sequence of previous queries and their answers, is larger than g. In our notation we can write this as the requirement that the following conditional expectation is larger than g almost everywhere:

$$\Pr_{\Delta}\left(E\left(\mathcal{G}(x_{i_{n+1}}|V(\langle\vec{X}_{I_n},c(\vec{X}_{I_n})\rangle)) \quad | \quad \vec{X}_{I_n},c(X_{I_n})\right) > g\right) = 1$$

In somewhat more intuitive terms, a uniform lower bound on the information means that for any version space that can be reached by **QBC** with non-zero probability, the expected information gain from the next query of **QBC** is larger than g. In Section 6 we shall prove uniform lower bounds on the information gain of **QBC** for some important learning problems.

We now give the theorem that relates the bound on the information gain of **QBC** to its expected prediction error.

THEOREM 1 *If a concept class* **C** *has VC-dimension* $0 < d < \infty$ *and the expected information gain of queries made by* **QBC** *is uniformly lower bounded by* $g > 0$ *bits, then the following holds with probability larger than* $1 - \delta$ *over the random choice of the target concept, the sequence of examples, and the choices made by* **QBC***:*

- *The number of calls to* **Sample** *that* **QBC** *makes is smaller than*

$$m_0 = \max\left(\frac{4d}{e\delta}, \frac{160(d+1)}{g\epsilon}\max\left(6, \ln\frac{80(d+1)}{\epsilon\delta^2 g}\right)^2\right) \quad . \tag{5}$$

- *The number of calls to* **Label** *that* **QBC** *makes is smaller than*

$$n_0 = \frac{10(d+1)}{g}\ln\frac{4m_0}{\delta} \quad ,$$

  In other words, it is an exponentially small fraction of the number of calls to **Sample**.[4]

- *The probability that the* **Gibbs** *prediction algorithm that uses the final version space of* **QBC** *makes a mistake in its prediction is smaller than* ϵ.

Before we proceed to prove the theorem, let us give a brief intuitive sketch of the argument (See Figure 3). The idea is that if a concept class is learnable then, after observing many labeled examples, the conditional distribution of the labels of new examples is highly biased to one of the two labels. This means that the information gained from knowing the label of a random example is small. This, in turn, means that the increase in the cumulative information from a sequence of random examples becomes slower and slower as the sequence gets longer. On the other hand, if the information gained from the queries of **QBC** is lower bounded by a constant, then the cumulative information gain from the sequence of queries increases linearly with the number of queries. It is clear that the information from the labels of the queries alone is smaller than the information from the labels of all the examples returned by **Sample**. The only way in which both rates of increase

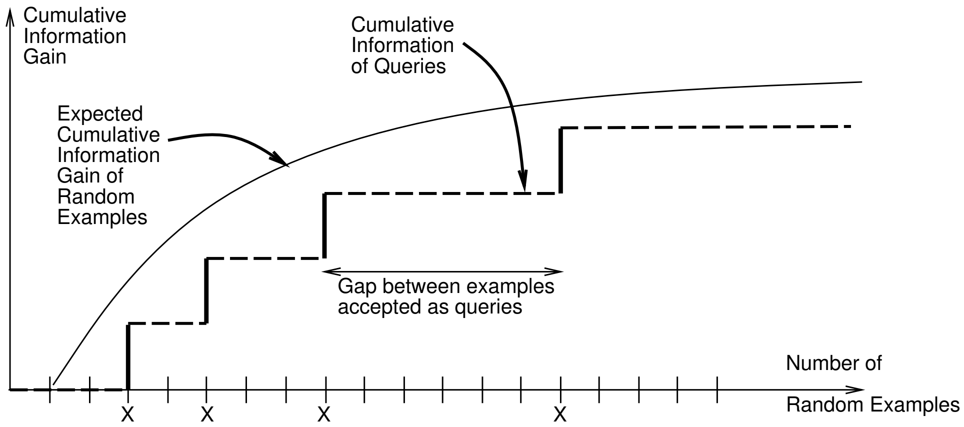

Figure 3. Each tag on the x axis denotes a random example in a specific typical sequence. The symbol X under a tag denotes the fact that the example was chosen as a query.

can hold without violating this simple inequality is if the number of examples that are rejected between consecutive queries increases with the number of queries. As a result the termination criterion of **QBC** will hold, and the algorithm will output its final prediction rule after a reasonably small number of queries. The prediction rule that is output is the Gibbs prediction rule, using the final version space that is defined by all the labeled examples seen so far. The probability of making a prediction error using this rule is, by definition, equal to the probability of a disagreement between a hypothesis that is randomly chosen according to the prior distribution restricted to the version space and a concept that is independently chosen according to the same distribution. This probability is also equal to the probability of accepting a random example as a query when using this version space. The termination condition is fulfilled only if a large number of random examples are not accepted as queries, which implies that the probability of accepting a query or making a prediction mistake when using the final version space is small. We shall prove the theorem using the following three lemmas.

LEMMA 1 *If the expected instantaneous information gain of the query algorithm is uniformly lower bounded by $g > 0$ bits, then*

$$Pr_{\Delta}(\mathcal{I}(\langle \vec{X}_{I_n}, c(\vec{X}_{I_n})\rangle) < \frac{g}{2}n) \leq e^{-\frac{g}{10}n} \tag{6}$$

Proof: The definition of a uniform lower bound on the expected information gain means that for any $n > 0$, for all sequence of of n queries $\langle \vec{X}_{I_n}, c(\vec{X}_{I_n})\rangle$, excluding possibly a set of measure zero, the expected information gain from the $n+1$st query is lower bounded by g. Put in another way, this means that the random variables

$$Y_i = \mathcal{I}(\langle \vec{X}_{I_i}, c(\vec{X}_{I_i})\rangle) - \mathcal{I}(\langle \vec{X}_{I_{i-1}}, c(\vec{X}_{I_{i-1}})\rangle) - g$$

form a sequence of sub-martingale differences. As the instantaneous information gain is bounded between 0 and 1, we get that $-g \leq Y_i \leq 1-g$. We can thus use Hoeffding's bound on the tails of bounded step sub-martingales (McDiarmid,1989) [5] from which we know that for any $\epsilon > 0$

$$\Pr(\sum_{i=1}^{n} Y_i \leq -\epsilon n) \leq \left[(\frac{g}{g+\epsilon})^{g+\epsilon} (\frac{1-g}{1-g-\epsilon})^{1-g-\epsilon} \right]^n .$$

Setting $\epsilon = \lambda g$ and taking logs we get

$$\begin{gathered} \Pr(\textstyle\sum_{i=1}^{n} Y_i \leq -\lambda g n) \leq \\ \exp\left(\left(-(1+\lambda) g \ln(1+\lambda) + (1-(1+\lambda)g) \ln \frac{1-g}{1-(1+\lambda)g} \right) n \right) \leq \\ \exp\left((\lambda - (1+\lambda)\ln(1+\lambda))\, gn \right) . \end{gathered}$$

Choosing $\lambda = 1/2$ we get the bound. ∎

LEMMA 2 *The probability that the predictions made by* **QBC** *are wrong (after its main loop has terminated) is smaller than ϵ with probability larger than $1-\delta/2$.*

Proof: Assume that the probability of a wrong prediction is larger than ϵ. As discussed in the informal part of the proof, this implies that the probability of accepting a random example as a query with the final version space, is also larger than ϵ. It thus remains to show that the probability that **QBC** stops when the probability of accepting a query is larger than ϵ is smaller than $\delta/2$.

The termination condition of **QBC** is that all t_n examples tested after the nth query are rejected. If the probability of accepting a random example is larger than ϵ then this probability is smaller than $(1-\epsilon)^{t_n}$. From the definition of t_n we get that

$$(1-\epsilon)^{\frac{1}{\epsilon} \ln \frac{\pi^2 (n+1)^2}{3\delta}} \leq e^{-\ln \frac{\pi^2(n+1)^2}{3\delta}} = \frac{3\delta}{\pi^2 (n+1)^2} .$$

Summing this probability over all possible values of n from zero to infinity we get the statement of the lemma. ∎

In (Haussler, Kearns & Schapire,1994) it was shown that if the VC-dimension of a concept class is d, then the expected information gain from m random examples is bounded by $(d+1)\log(m/d)$. Here we show that the probability that the information gain is much larger than that is very small.

LEMMA 3 *Assume a concept c is chosen at random from a concept class with VC dimension d. Fix a sequence of examples $\vec{X}$, recall that $\vec{X}_M$ denotes the first m examples. Then*

$$Pr_{c \in \mathcal{P}} \left(\mathcal{I}(\langle \vec{X}_M, c(\vec{X}_M) \rangle) \geq (d+1)(\log \frac{em}{d}) \right) \leq \frac{d}{em} . \tag{7}$$

Proof: From Sauer's Lemma (Sauer,1972) we know that the number of different labelings created by m examples is at most $\sum_{i=0}^{d}\binom{m}{i} \le (em/d)^d$. The expected cumulative information gain is equal to the entropy (base 2) of the distribution of the labels and is maximized when all the possible labelings have equal probability. This gives an upper bound of $d \log \frac{em}{d}$ on the expected cumulative information gain. Labelings that have cumulative information gain larger by a than this expected value, must have probability that is smaller by 2^a than the labels in the equipartition case. As the number of possible labelings remains the same, the total probability of all concepts that give rise to such labelings is at most 2^{-a}. Choosing $a = \log \frac{em}{d}$ we get the bound. ∎

Proof of Theorem 1: We consider a randomly chosen element of the event space $\langle c, \vec{X}, I \rangle$. Our analysis involves the first m_0 random examples presented to **QBC**, $\vec{X}_{M_0}$, and the first n_0 queries that **QBC** would filter if it never halts, $\vec{X}_{I_{n_0}}$. We denote the number of queries that **QBC** makes during the first m_0 examples by n, i.e. $n = |I \cap M_0|$. The claim of the theorem is that, with probability at least $1-\delta$, the algorithm halts before testing the $m+1$st example, the number of queries it makes, n, is smaller than n_0, and the hypothesis it outputs upon halting has error smaller than ϵ. We shall enumerate a list of conditions that guarantee that all of these events occur for a particular random choice of examples and of internal randomization in **QBC**. By showing that the probability of each of those conditions to fail is small we get the statement of the theorem.

The conditions are:

1. <u>The cumulative information content of the first n_0 queries is at least $gn_0/2$.</u>
   From Lemma 1 we get that in order for this condition to hold with probability larger than $1-\delta/4$ it is sufficient to require that

$$n_0 \ge \frac{10}{g} \ln \frac{4}{\delta} \ . \tag{8}$$

2. <u>The cumulative information content from the first m_0 examples is at most $(d+1)(\log \frac{em_0}{d})$.</u>
   From Lemma 3 we get that in order for this condition to hold with probability larger than $1-\delta/4$ it is sufficient to require that

$$m_0 \ge \frac{4d}{e\delta} \ . \tag{9}$$

3. <u>The number of queries made during the first m_0 examples, n, is smaller than n_0.</u>
   The condition follows from conditions 1 and 2 if

$$\mathcal{I}(\langle \vec{X}_{I_{n_0}}, c(\vec{X}_{I_{n_0}}) \rangle) \ge \mathcal{I}(\langle \vec{X}_{M_0}, c(\vec{X}_{M_0}) \rangle) \tag{10}$$

This is because if $n \geq n_0$ then the information gained from the queries asked during the first m_0 examples is larger than the total information gained from the m_0 examples, which is impossible. In order for (10) to hold, it is sufficient to require that

$$n_0 > \frac{2(d+1)}{g}(\log \frac{em_0}{d}) \ . \tag{11}$$

4. <u>The number of consecutive rejected examples guarantees that the algorithm stops before testing the $m_0 + 1$st example.</u>
Notice that the threshold t_i increases with i. Thus if at least t_n consecutive examples from among the first m_0 examples are rejected, the algorithm is guaranteed to halt before reaching the $m_0 + 1$st example. As there are $m_0 - n$ rejected examples, the length of the shortest run of rejected examples is at least $(m_0 - n)/(n+1)$. We require that this expression is larger than t_n, and use the fact that condition 3 holds, i.e. that $n < n_0$. Using these facts it is sufficient to require that

$$m_0 \geq \frac{2(n_0+1)}{\epsilon} \ln \left[\frac{\pi^2}{3\delta}(n_0+1)^2 \right] \ . \tag{12}$$

5. <u>The Gibbs prediction hypothesis that is output by the **QBC** has probability smaller than ϵ of making a mistaken prediction.</u>
From Lemma 2 we get that the probability of this to happen is smaller than $\delta/2$.

We see that if Equations (8), (9), (11), and (12) hold, then the probability that any of the four conditions fails is smaller than δ.It thus remains to be shown that our choices of n_0 and m_0 guarantee that these equations hold. Combining Equations (8) and (11), we get that it is sufficient to require that $m_0 \geq 2$, $d \geq 1$, and

$$n_0 + 1 = \frac{10(d+1)}{g} \ln \frac{4m_0}{\delta} \tag{13}$$

Plugging this choice of n_0 into Equation (12), we get the following requirement on m_0:

$$m_0 \geq \frac{40(d+1)}{\epsilon g} \ln \frac{4m_0}{\delta} \ln \left[\frac{20(d+1)}{\delta g} \ln \frac{4m_0}{\delta} \right] \ . \tag{14}$$

It is simple algebra to check that the following choice of m_0 and satisfies Equations (9) and (14):

$$m_0 = \max \left(\frac{4d}{e\delta}, \frac{160(d+1)}{g\epsilon} \max \left(6, \ln \frac{80(d+1)}{\epsilon \delta^2 g} \right)^2 \right) \ , \tag{15}$$

Equations (13) and (15) guarantee that the conditions 1-5 hold with probability at least $1 - \delta$. ■

6. Concept classes that are efficiently learnable using QBC

According to Theorem (1) above, if query by committee yields high information gain, then it yields a rapidly decreasing generalization error. Here we discuss some geometric concept classes for which a uniform lower bound on the information gain exists, and hence to which the theorem is applicable.

Our main analysis is for a learning problem in which concepts are intersections of half-spaces with a compact and convex subset of R^d. In this case the concept class itself can be represented as a compact and convex subset of R^d and each example partitions the concept class by a $d-1$ dimensional hyperplane. In Section 6.1, we sketch a proof of a uniform lower bound on the information gain of **QBC** that does not depend on the dimension d, for the case in which both $\mathcal{D}$ and $\mathcal{P}$ are uniform. The proof, which is detailed in Appendix A. is based on a variational analysis of the geometry of the version space. In Section 6.2 this result is extended to the case of non-uniform input distribution and prior and applied to the perceptron learning problem.

6.1. Uniformly distributed half-spaces

In this subsection we prove a lower bound on the information gain for a simple geometric learning problem to which we shall refer as the "parallel planes" learning problem.

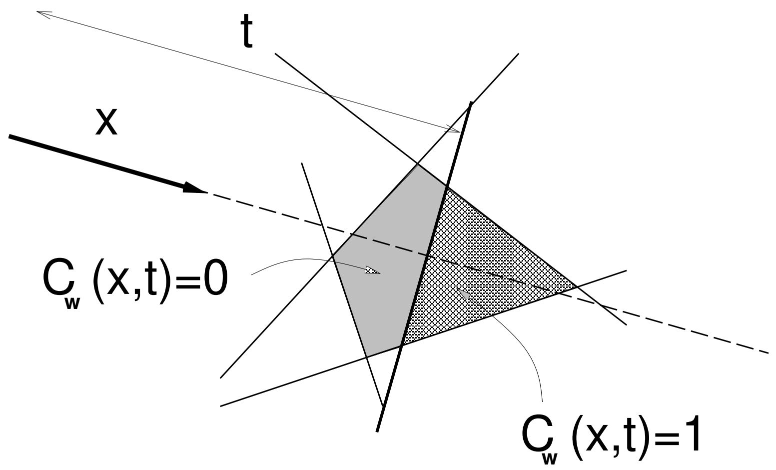

Figure 4. A figure of the two dimensional concept class defined by Equation (16) for $d=2$. The shaded area corresponds to a typical convex version space V which is defined by a set of half spaces corresponding to several examples. This version space is bisected by a new unlabeled example defined by $\vec{x}$ and t.

We define the domain, X, to be the set of all pairs of the form $(\vec{x}, t)$, where $\vec{x}$ is a vector in R^d whose length is 1, which we refer to as the "direction" of the example, and t is a real number in the range $[-1, +1]$, to which we refer as the offset (see Figure 6.1). In other words $X = S^d \times [-1, +1]$, where S^d denotes the unit sphere around the origin of R^d. In this section we assume that the distribution $\mathcal{D}$ on X is uniform.[6] The concept class, $\mathbf{C}$, is defined to be a set of binary functions over X, parameterized by vectors $\vec{w} \in R^d$, $||\vec{w}||_2 \leq 1$, that are defined as follows

$$c_{\vec{w}}(\vec{x}, t) = \begin{cases} 1, & \vec{w} \cdot \vec{x} \geq t \ , \\ 0, & \vec{w} \cdot \vec{x} < t \end{cases} . \tag{16}$$

We assume that the prior distribution is uniform on B^d—the unit ball of radius one around the origin. This concept class is very similar to the class defined by the perceptron with variable threshold. [7] However, note that in this case the threshold, t, is part of the input, and not a parameter that defines the concept. This concept class is a bit strange, but as we shall see, the results we can prove for it can be extended to more natural concept classes such as the perceptron.

The information gain from random examples vanishes as d goes to infinity. The reason for this is that in high dimension, the volume of the sphere is concentrated near the equator. A typical random example will cut the sphere some distance away from the equator, in which case the sphere will fall into two pieces of very unequal volume. The piece containing the equator will contain almost all of the volume. This geometric example illustrates why query algorithms are especially important in high dimensions. Query by committee solves this problem by choosing two random points in the sphere. Since these two points are likely to be near the equator, an example that separates them is likely to be near the equator. For this reason, query by committee can attain an information gain that remains finite in high dimensions.

In our proof of the uniform lower bound on the expected information gain of **QBC** we use two properties of the version spaces for this concept class. The first property is that each example $(\vec{x}, t)$ cuts the version space by a plane that is orthogonal to the direction $\vec{x}$ and has offset t from the origin.[8] As t is uniformly distributed, the planes that cut the version space in any fixed direction have a uniformly distributed offset that spans the width of the version space in that direction. The second property is that all version spaces that can be generated when learning this concept class are bounded convex sets because they are defined as the intersection of a ball with a number of half-spaces.

As discussed in Section 4, both the expected information gain of an example and the probability that the example is accepted by **QBC** are quantities that depend on the ratio between the probabilities of the two parts of the version space that are created by the example. Based on these observations we can reduce our problem to a one dimensional problem. Fix a particular direction $\vec{x}$. Let $F_{\vec{x}} : [-1, +1] \to [0, 1]$ be the fraction of the version space, V, that is on one side of the plane defined by $\vec{x}$ and t, i.e.

$$F_{\vec{x}}(t) = \frac{\Pr_{c_{\vec{w}} \in \mathcal{P}} \left(c_{\vec{w}} \in V | c_{\vec{w}}(\vec{x}, t) = 0\right)}{\Pr_{c_{\vec{w}} \in \mathcal{P}} \left(c_{\vec{w}} \in V\right)} . \tag{17}$$

We call F the *volume function* of the version space. The probability that **QBC** accepts the example $(\vec{x}, t)$ is $2F_{\vec{x}}(t)(1-F_{\vec{x}}(t))$, and the expected information gain from the example is $\mathcal{H}(F_{\vec{x}}(t))$. As t is uniformly distributed, the expected information gain from the examples whose direction is $\vec{x}$ is

$$G(F_{\vec{x}}) = \frac{\int_{-1}^{+1} F_{\vec{x}}(t)(1 - F_{\vec{x}}(t))\mathcal{H}(F_{\vec{x}}(t))\, dt}{\int_{-1}^{+1} F_{\vec{x}}(t)(1 - F_{\vec{x}}(t))\, dt} \; . \tag{18}$$

Our result is a lower bound on the value of $G(F_{\vec{x}})$. The proof is based on finding the convex version space that produces the smallest value of $G(F_{\vec{x}})$. This body is constructed of two isomorphic cones connected at their bases, we call this body a "two-cone". Barland (Barland,1992, Theorem5), analyzes a similar problem. He finds the convex body that achieves the minimal value of the functional $\int_{-1}^{+1} \min(F_{\vec{x}}(t), 1 - F_{\vec{x}}(t))\, dt$. The analysis of the minimum for this functional is much simpler, interestingly, Barland finds that the body which achieves the minimum is the same as the one which achieves the minimum of the functional G.

THEOREM 2 *The functional $G(F_{\vec{x}})$, defined for volume functions of convex bodies in R^d, assumes a unique minimum at the two-cones body defined above. The value of G at the minimum is at least* $1/9 + 7/(18 \ln 2) > 0.672$ *bits, for any dimension* d.

This theorem gives us a lower bound on the expected information gain of a single query of **QBC** for the "parallel planes" learning problem defined at the beginning of this section. In Section 6.2 we shall use this theorem to prove that **QBC** is an effective query algorithm for learning perceptrons.

Proof: Here we give the main part of the proof. The more technical details are formulated in Lemmas 4, 5, 6 and 7, whose proofs are given in appendix A.

The proof is based on a variational analysis of the functional G. We shall show that the volume function that corresponds to "two-cones" minimizes this functional. We shall show that any other volume function of a convex body can be slightly altered in a way which decreases the value of G and maintains the correspondence with some convex body.

We shall bound the value of $G(F_{\vec{x}})$ independently of the direction $\vec{x}$. Our bound depends only on the fact that the version space is a bounded convex set in R^n and that the distribution in it is uniform. We thus drop the subscript $\vec{x}$ from $F_{\vec{x}}(\cdot)$. As $F(-1) = 0$, $F(+1) = 1$, and $\mathcal{H}(1) = \mathcal{H}(0) = 0$, we will, without loss of generality, extend the definition of $F(t)$ to all of R by defining it to be zero for $t \leq -1$ and one for $t \geq 1$. We then redefine the integrals in the definition of $G(F)$ in Equation (18) to be from $-\infty$ to ∞. It is easy to check that $G(F(t)) = G(F(at + b))$ for any $a, b \neq 0$. Thus, without loss of generality, the support of the volume function is $[-1, +1]$ and $F(0) = 1/2$.

Consider the right half of the body, i.e. the set of points whose t coordinate is at least 0. Take the union of this half with its symmetric reflection at the plane $t = 0$. Similarly, generate a symmetric body from the left side of the original body. The two resulting bodies are reflection symmetric but usually not convex. Their volume functions are:

$$F_-(t) = \begin{cases} F(t) & t \le 0 \\ 1 - F(-t) & t > 0 \end{cases},$$

$$F_+(t) = \begin{cases} 1 - F(-t) & t \le 0 \\ F(t) & t > 0 \end{cases},$$

It is easy to see that either $G(F_+) \le G(F)$ or $G(F_-) \le G(F)$. Thus, in order to prove a lower bound on $G(F)$ for all convex bodies, it is sufficient to prove a lower bound for volume functions that correspond to reflection-symmetric bodies for which each half is convex. Our variational manipulations will apply to one half of the symmetric body (say $t \ge 0$) and carry over by reflection to other half. As we shall show, the minimum for one half is obtained for a cone with a base at $t = 0$. Its symmetric reflection, the two-cone body, happens to be a convex body. Thus the two cone body gives the minimum of $G(F)$ for all convex bodies.

Our goal is thus to find the a volume function $F : [0, \infty) \to [1/2, 1]$ of the right half of a convex body, which minimizes the functional

$$G(F) = \frac{\int_0^\infty F(x)(1 - F(x))\mathcal{H}(F(x))\,dx}{\int_0^\infty F(x)(1 - F(x))\,dx}\ . \tag{19}$$

We find it convenient to define the functions $K(t) = F(t)(1 - F(t))$, and $Q(x) = \mathcal{H}(1/2 - \sqrt{1 - 4x}/2)$. It is easy to verify that $\mathcal{H}(F) = Q(K)$, and that Equation (19) can be written as

$$G(K) = \frac{\int_0^{+\infty} K(t)Q(K(t))\,dt}{\int_0^{+\infty} K(t)\,dt}\ . \tag{20}$$

The changes in $G(K)$ that are induced by small changes in the function K can be approximated by a linear functional, called the Fréchet derivative,[9] as follows

$$G(K + \Psi) = G(K) + \int_0^{+\infty} \nabla G[K](t)\Psi(t)\,dt + o\left(\int_0^{+\infty} \Psi(t)^2\,dt\right)\ .$$

The Fréchet derivative $\nabla G[K]$ is a function from $[0, \infty)$ into R and $\nabla G[K](t)$ is the value of this function at the point t. The derivative is calculated by formally differentiating the functional $\nabla G[K]$ with respect to $K(t)$. Thus

$$\begin{aligned} \nabla G(t) &= \\ &\frac{\int_0^{+\infty} K(s)\,ds\ \frac{\partial}{\partial K(t)}\left(K(t)Q(K(t))\right) - \int_0^{+\infty} K(s)Q(K(s))\,ds\ \frac{\partial}{\partial K(t)}K(t)}{\left(\int_0^{+\infty} K(s)\,ds\right)^2} \qquad (21) \\ &= \frac{1}{\int_0^{+\infty} K(s)\,ds}\left[Q(K(t)) + K(t)\frac{\partial}{\partial K(t)}Q(K(t)) - G(K)\right] \end{aligned}$$

We first consider the behavior of the sum of the first two terms in the square brackets. Denote $K(t)$ by y. A direct calculation shows that $Q(y) + y\frac{\partial}{\partial y}Q(y)$ is a strictly increasing function of y in the range $0 \leq y \leq 1/4$, which is the range of $K(t)$. It is 0 for $y = 0$ and 1 for $y = 1/4$.

As $0 \leq G(K) \leq 1$ the third term is in the range of the sum of the first two terms. As $K(t)$ is decreasing for positive t, it follows that there is some point $w > 0$, which is a function of K, such that for all $0 \leq t \leq w$, $\frac{\partial}{\partial K(t)}G(K(t)) > 0$, and for all $t > w$, $\frac{\partial}{\partial K(t)}G(K(t)) < 0$. The parameter w is of critical importance in the rest of the paper, and we shall refer to it is the "pivot point". In terms of the volume function F, for $t > 0$, F increases when K decreases and vice versa. Thus if the variation $\Psi(t)$ is non-negative for points below the pivot point, non-positive for points above the pivot point, and $\int_0^{+\infty} \Psi(t)^2\, dt$ is sufficiently small then $G(K(t) + \Psi(t)) < 0$ as desired.

We shall construct suitable variations in proving lemma 5. For now, let B be the convex body whose volume function is $F(t)$. Consider the functions $f(t)$ and $r(t)$ defined as follows:

$$f(t) = \frac{dF(t)}{dt}\ ; \qquad r(t) = \sqrt[d-1]{\frac{f(t)}{c_{d-1}}}\ ,$$

where $c_d - 1$ is the volume of the $d-1$ dimensional unit ball. The function $F(t)$ is equal to the total volume of the body B in the range $(-\infty, t]$, so $f(t)$ is the $d-1$ dimensional volume of the *slice* of B at t. We call $r(t)$ the *radius function* because if $\tilde{B}$ is a body of revolution obtained by rotating (the planar graph of) the function $t \rightarrow r(t)$ around the axis $r(t) = 0$, then the volume functions that correspond to $\tilde{B}$ and to B are the same. Moreover,the following Lemma characterizes radius functions of convex bodies

LEMMA 4 *1. The radius function of any convex body is concave.*

2. The body of revolution that is generated by a concave radius function is convex.

The proof of the lemma is given in Appendix A. Thus the search for the minimum of $G(K)$ over convex bodies (for $t \geq 0$) can be restricted to bodies of revolution created by rotating a concave radius function $r(t)$.

The proof of the theorem is concluded by proving the following lemmas, the details are in Appendix A.

LEMMA 5 *If the convex body with volume function F is not a cone with base at the hyperplane $t = 0$ then there exists an admissible variation Ψ such that $G(F + \Psi) < G(F)$.*

LEMMA 6 *The minimum of G over convex bodies is achieved.*

From Lemmas 5 and 6 it follows that the minimum of $G(F)$ is achieved for the two-cone body. Finally a simple calculation gives that

LEMMA 7 *The value of $G(F)$ for a two-cone body in R^d is at least* $1/9+7/(18\ln 2) > 0.672$ *for any dimension d.*

This concludes the proof of Theorem 2. ■

6.2. Perceptrons

In this section we apply Theorem 2 to the problem of learning perceptrons. The perceptron concept class is defined as the following set of binary functions over the unit ball

$$c_{\vec{w}}(\vec{x}) \begin{cases} 1, & \vec{x}\cdot\vec{w} \geq 0 \\ 0, & \text{otherwise} \end{cases}, \tag{22}$$

where $\vec{w}, \vec{x} \in R^d$, $||\vec{w}||_2 = 1$ and $||\vec{x}||_2 \leq 1$. The prior distributions are within some constants from the uniform distributions over the respective sets. As each $\vec{w}$ is a point on the surface of a d dimensional sphere, the initial version space is isomorphic to the unit sphere.

The section is organized as follows. We start by stating an extension of Theorem 2. We then discuss a technical issue regarding an initial phase of the learning procedure that is required in order to make the theorems apply. We then prove the main result of this section, which shows that, under some mild assumptions, the prediction error of the **QBC** algorithm, when learning decreases exponentially fast with the number of queries asked.

Theorem 2 can be generalized to cases where the prior and input distributions are not exactly uniform. We use the following definition

Definition. We say that a density $\mathcal{D}'$ is *within* λ of $\mathcal{D}$ if for every measurable set A, we have that $\lambda \leq \Pr_{\mathcal{D}}(A)/\Pr_{\mathcal{D}'}(A) \leq 1/\lambda$.

Using this definition, we get the following extension of Theorem 2:

THEOREM 3 *The value of the functional $G(F)$ for the parallel planes learning problem, when the prior distribution is within $\lambda_{\mathcal{P}}$ of uniform and the input distribution is within $\lambda_{\mathcal{D}}$ of uniform, is at least $\lambda_{\mathcal{P}}^4\lambda_{\mathcal{D}}(1/9 + 7/(18\ln 2)) > 0.672\lambda_{\mathcal{P}}^4\lambda_{\mathcal{D}}$ bits, independent of the dimension d.*

The proof is in Appendix B.

Using Theorem 3, we can prove that **QBC** is an efficient query algorithm for the perceptron concept class when the prior distribution and the distribution of examples are both close to uniform. We shall prove that there exists a lower bound on the information gain of the queries of **QBC**. However, our proof technique requires that the initial version space is not the complete unit sphere, but is restricted to be within a cone. In other words, there has to exist a unit vector $\vec{w}_0$ such that for any $\vec{w} \in V_0$ the dot product $\vec{w}\cdot\vec{w}_0$ is larger than some constant $\alpha > 0$.

This condition is annoying. However, it is not hard to guarantee that this condition holds by using an initial learning phase, prior to the use of **QBC**, that does not use filtering but rather queries on all the random instances supplied by **Sample**. Using the results of Blumer et al. we can bound the number of training examples that are needed to guarantee that the prediction error of an arbitrary consistent hypothesis is small (with high probability). As the distribution of the instances is close to uniform, a small prediction error implies that the hypothesis vector is within a small angle of the vector that corresponds to the target concept. The details of this argument are given in the following lemma.

LEMMA 8 *Assume that the distribution of the instances $\mathcal{D}$ is within $\lambda_{\mathcal{D}}$ from the uniform distribution in the unit ball. Suppose m random instances are chosen according to $\mathcal{D}$, labeled according to $f_{\vec{w}_0}(\cdot)$ and used to find a hypothesis $f_{\vec{w}}(\cdot)$ that is consistent with all the labeled instances.*

If

$$m \geq \max\left(\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{13}{\epsilon}\right) \quad \textit{where } \epsilon = \lambda_{\mathcal{D}}\cos^{-1}(\alpha)$$

then, with probability $1-\delta$ over the choice of the m random instances, $\vec{w}\cdot\vec{w}_0 \geq \alpha$.

Proof: If $\vec{w}\cdot\vec{w}_0 < \alpha$ then the angle between $\vec{w}$ and $\vec{w}_0$ is larger than $\cos^{-1}(\alpha)$. The examples on which $f_{\vec{w}}(\vec{x})$ is incorrect are those vectors in the unit ball for which $\vec{x}\cdot\vec{w} \geq 0$ and $\vec{x}\cdot\vec{w}_0 < 0$, or $\vec{x}\cdot\vec{w} < 0$ and $\vec{x}\cdot\vec{w}_0 \geq 0$. This defines a subset of the unit ball, constructed of two wedges, whose volume is at least $\cos^{-1}(\alpha)$ of the volume of the ball. As the distribution of the instances is within $\lambda_{\mathcal{D}}$ from the uniform distribution, the probability of this set is at least $\lambda_{\mathcal{D}}\cos^{-1}(\alpha)$.

On the other hand, as the VC dimension of the d dimensional perceptron is d we can use the classical uniform convergence bounds from (Blumer et al.,1989). Theorem 2.1 in (Blumer et al.,1989) guarantees that a hypothesis that is consistent with m labeled examples, chosen independently at random from an *arbitrary* distribution, has error smaller than ϵ with probability $1-\delta$ if

$$m \geq \max\left(\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{13}{\epsilon}\right) .$$

Combining these two arguments, we get the statement of the theorem. ∎

Assuming that an initial phase of learning from unfiltered instances is used to guarantee a bound on the maximal angle between vectors, we get the following theorem.

THEOREM 4 *For any $\alpha > 0$, let $\mathbf{C}_\alpha$ be the d dimensional perceptron concept class as defined in Equation (22), restricted to those concepts $c_{\vec{w}}$, such that $\vec{w}_0\cdot\vec{w} > \alpha$ for some unit vector $\vec{w}_0$. Let the prior distribution over $\mathbf{C}_\alpha$ be within $\lambda_{\mathcal{P}}$ of uniform and the input distribution be within $\lambda_{\mathcal{D}}$ from uniform. Then the expected information gain of the queries of* **QBC** *is larger than $0.672\alpha^{5d}\lambda_{\mathcal{P}}^4\lambda_{\mathcal{D}}$*

Proof: The version space for the perceptron is a region on the d-dimensional unit sphere that is bounded by a set of great circles. We shall transform this problem into a special case of the parallel planes learning problem defined in Section 6.1.

Because we assume the existence of the vector $\vec{w}_0$ we can define a one-to-one mapping of the version space to a bounded convex subset of R^{d-1}. We can assume, without loss of generality, that $\vec{w}_0 = \{1, 0, \ldots, 0\}$. We can also assume that $||\vec{x}||_2 = 1$, because all instances $\vec{x}$ whose length is smaller than 1 can be mapped to $\vec{x}/||\vec{x}||_2$ without changing the label assigned to them by the concepts. The distribution over the surface of the unit sphere that is created in this way is within $\lambda_{\mathcal{D}}$ of uniform.

In this case the mapping of the concepts is defined by transforming the vector $\vec{w} = \{w_1, w_2, \ldots, w_d\}$ that lies on the unit sphere to the $d-1$ dimensional vector $\vec{w}' = \{w_2/w_1, w_3/w_1, \ldots, w_d/w_1\}$. The corresponding mapping of the instances maps the instance $\vec{x} = \{x_1, \ldots, x_d\}$ that lies on the unit sphere to the pair $\vec{x}' = \{x_2, \ldots, x_d\}/\sqrt{\sum_{i=2}^d x_i^2}$ and $t = -x_1/\sqrt{\sum_{i=2}^d x_i^2}$. It is easy to see that the condition that defines the perceptron $\vec{w} \cdot \vec{x} \geq 0$ is equivalent to $\vec{x}' \cdot \vec{w}' \geq t$, which is the condition that defines the corresponding parallel-plane concept.

The condition $\vec{w} \cdot \vec{w}_0 > \alpha$ is, in this case, equivalent to $w_1 > \alpha$. It is easy to check that the only examples in the transformed concept space that can be labeled both 0 and 1 by some concept in $\mathbf{C}_\alpha$ are those for which $\sqrt{\sum_{i=2}^d x_i^2} > \alpha$. This implies that the increase in the volume of an infinitesimal part of the instance space is by a factor of at most α^{-d}. Thus as the distribution over the instances on the surface of the unit sphere is within $\lambda_{\mathcal{D}}$ of uniform, the distribution over the transformed instance space is within $\alpha^d \lambda_{\mathcal{D}}$ of uniform.

To bound the distance of the prior distribution from uniform, consider the mapping of an infinitesimally small region of the version space from the sphere to the plane. Figure 5 illustrates this transformation for a two dimensional perceptron. This transformation maps the hyperspherical region to a larger region in the hyperplane. The factor by which the volume is increased is between 1 and α^{-d}. This can be seen by separating the transformation into two steps. In the first step, the region on the unit hypersphere is mapped to a region on a larger hypersphere. The radius of this larger hypersphere is at most α^{-1}, thus the increase in the volume is by a factor of at most $\alpha^{-(d-1)}$. In the second step, the region on the large hypersphere is mapped to the hyperplane, as the region is infinitesimally small, it can be approximated by a linear region. The increase in the volume of the region in this step is by a factor of α^{-1}. Multiplying the two factors we get α^{-d}.

As the prior distribution over the sphere is within $\lambda_{\mathcal{P}}$ of uniform, the distribution over the hyperplane that is generated by the mapping is within $\lambda_{\mathcal{P}} \alpha^d$ of uniform.

We thus have a special case of the parallel plane learning problem with close to uniform distributions. Using Theorem 3, we get the statement of the theorem.

■

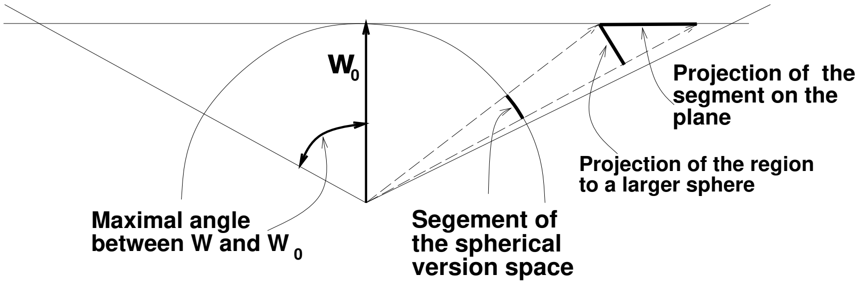

Figure 5. The transformation that maps the spherical version space unto the hyperplane.

6.3. Using an incorrect prior distribution

Up to this point we have made the assumption that the learning algorithm is using the correct prior distribution on the concept space $\mathcal{P}$. In this section we show how this assumption can be weakened.

Definition. [10] We say that a distribution $\mathcal{P}$ is λ-dominated by a distribution $\mathcal{P}$ if, for any event A, $\Pr_{\mathcal{P}}(A) \leq \lambda \Pr_{\mathcal{P}'}(A)$.

Suppose that **QBC** uses a distribution $\mathcal{P}'$ that λ-dominates $\mathcal{P}$ for some $0 < \lambda < \infty$ such that there is a uniform lower bound on the expected information gain of **QBC** with respect to $\mathcal{P}'$. The following theorem replaces Theorem 1 for this case.

THEOREM 5 *If a concept class* **C** *has VC-dimension* $0 < d < \infty$ *and the expected information gain of queries made by* **QBC** *when using the prior* $\mathcal{P}'$ *is uniformly lower bounded by* $g > 0$ *bits, and if* $\mathcal{P}$ *is* λ_c*-dominated by* $\mathcal{P}'$ *for some* $0 < \lambda_c < \infty$ *then the following holds with probability larger than* $1 - \delta$ *over the random choice of the target concept (with respect to* $\mathcal{P}$*), the sequence of examples, and the choices made by* **QBC***:*

- *The number of calls to* **Sample** *that* **QBC** *makes is smaller than*

$$m_0 = \max\left(\frac{4d}{e\delta}, \frac{160(d+1)}{g\lambda_c^2\epsilon}\max\left(6, \ln\frac{80(d+1)}{\lambda_c^2\epsilon\delta^2 g}\right)^2\right) . \tag{23}$$

- *The number of calls to* **Label** *that* **QBC** *makes is smaller than*

$$n_0 = \frac{10(d+1)}{g}\ln\frac{4m_0}{\delta} ,$$

In other words, it is an exponentially small fraction of the number of calls to **Sample***.*

- *The probability that the* **Gibbs** *prediction algorithm that uses the final version space of* **QBC** *makes a mistake in its prediction is smaller than* ϵ.

Note that while the number of calls to **Sample** increases by about a factor of λ_c^2, the number of queries increases only by an additive term of about $2 \log \lambda_c$.

Sketch of proof: It is clear that the arguments given in the proofs of Lemmas 1 - 3 and Theorem 1 hold if $\mathcal{P}$ is replaced by $\mathcal{P}'$ throughout. This implies that, with high probability, the error of a **Gibbs** prediction algorithm that uses the final version space of **QBC** is smaller than ϵ', or

$$E_{\mathcal{D}} \left[\Pr_{c \sim \mathcal{P}', h \sim \mathcal{P}'} \left[c(x) \neq h(x) \right] \right] \leq \epsilon' \ .$$

The assumption that $\mathcal{P}$ is λ_c-dominated by $\mathcal{P}'$ implies that

$$E_{\mathcal{D}} \left[\Pr_{c \sim \mathcal{P}, h \sim \mathcal{P}} \left[c(x) \neq h(x) \right] \right] \leq \lambda_c^2 \epsilon' \ .$$

By increasing m_0 by a factor of λ_c^2 we get that $\lambda_c^2 \epsilon' = \epsilon$, from which the statement of the theorem follows. ■

7. Learning using unlabeled examples and membership queries

The **QBC** algorithm uses unlabeled examples in order to reduce the number of labeled examples that it needs to know. While **QBC** is a very simple algorithm it is not the only way of using the information provided by random unlabeled examples. In this section we make the observation that in the learning framework defined in this paper there is a general scheme for query filtering. This scheme is potentially more computationally intensive than **QBC**, however, it is applicable in more generality than **QBC**.

The main observation is that the oracles **Sample** and **Gibbs**, defined in Section 2 allow the learning algorithm to estimate the expected error of any prediction rule. In this way the algorithm can calculate the expected improvement of making any particular query.

The prediction rule used by **QBC** is to select a random consistent hypothesis h using **Gibbs**, and then label the instance with $h(x)$. In general, any prediction rule defines a conditional distribution of the label given the instance. The error of a prediction rule for a given instance $x \in X$ and concept $c \in C$ is the probability that the prediction assigns to the incorrect label $1 - c(x)$. The expected error of the prediction rule is defined by selecting x at random according to $\mathcal{D}$ and a c at random according to $\mathcal{P}$. The oracles **Sample** and **Gibbs** generate random selections from $\mathcal{D}$ and $\mathcal{P}$ respectively. Thus, disregarding computational complexity, we can approximate the expected error of any prediction rule using sufficiently large samples of instances and hypotheses.

The dependence of the prediction rule generated by **QBC** on the labeled instances seen in the past is defined via the version space V. In general, any learning algorithm defines a mapping from sets of labeled instances to prediction rules. The estimate of the error of a prediction rule thus defines a measure of the quality of a set of labeled examples. If we are given an *unlabeled* instance, we can estimate the distribution of the label of the instance by using **Gibbs**. In this way we can estimate the expected reduction in the prediction error that will result from knowing the correct label of any instance. A reasonable heuristic for filtering queries is to select those instances that cause the largest reduction in the prediction error. If after observing any set of labeled instances the learning algorithm can find an instance which reduces the expected prediction error by a constant multiplicative factor, then the prediction error decreases exponentially fast in the number of queries asked. Of course, instances that cause such a reduction might not always exist, and even if they exist, the problem of finding them efficiently is potentially hard.

The algorithm analyzed in this paper, **QBC**, is an efficient variant of this heuristic. The general heuristic described above makes a large number of calls to the oracles **Sample** and **Gibbs**, algorithm **QBC** makes much fewer calls. More specifically, the dependence of the number of calls to **Sample** on the desired error, ϵ is[11] $\tilde{O}(1/\epsilon)$, which is the same dependence achieved by the algorithm that makes a query on each instance that it gets from **Sample**. The algorithm makes twice as many calls to **Gibbs** as it makes to **Sample**. It is not clear if this is close to optimal, however, it is certainly much smaller than the number of calls that is suggested in the heuristic described above. The exponential decrease of the error of **QBC** as a function of the number of queries has been established for a restricted family of parameterized concept classes. Establishing the effectiveness of **QBC** for more general concept classes or proving that it will not be effective for general families of concept classes is an interesting open problem.

While the general heuristic described in this section is not efficient, it is applicable in much more general situations than **QBC**. For example, the outcomes do not have to be binary or even discrete, and the relation between them and the inputs can be stochastic rather than deterministic. Finding learning algorithms that learn efficiently in this more general frameworks is another interesting open problem.

8. Summary

We have proved that the Query by Committee algorithm is an efficient query algorithm for the perceptron concept class with distributions that are close to uniform. This gives a rigorous proof to the results given in (Seung, Opper & Sompolinsky,1992) which were obtained using the replica method of statistical mechanics. It also generalizes their results by relaxing the requirements on the distribution of the examples and on the prior distribution. In addition, we show that exact knowledge of the prior distribution is not required. It is sufficient if the ratio between the assumed prior and actual prior is bounded by a constant factor.

We have proved that, in general, if the queries that are filtered by the query by committee algorithm have high expected information gain then the prediction error is guaranteed to decrease rapidly with the number of queries. By proving that this is the case for the perceptron learning problem, we have achieved our main result.

We hope that lower bounds on the expected information gain of **QBC** can be proven for other concept classes. It seems that it would be very useful, in this context, to generalize Theorem 1 to allow cases in which the expected information gain is small to occur with some small probability.

There are several issues that we do not discuss in this paper. First, one would like to know whether the results can be extended to concept classes other than perceptrons. Second, it is of great practical importance to analyze more general scenarios. In the "noisy" case, the learner sometimes observes a corrupted label, which is different from the correct label associated with the instance. A related case is the "probabilistic" case, in which the relationship between the instance and the label is described by a conditional distribution. An even more general case is the "agnostic" scenario, in which the only assumption is that there is some joint distribution over instances and labels from which examples are drawn independently at random. Extending our analysis to any of these more general cases is an an open problem which is important for making the analysis more relevant to practical applications.

Though theoretical results for such models are lacking, there is empirical evidence that extensions of the **QBC**algorithm can be used to learn noisy and probabilistic models, such as hidden Markov models (Dagan & Engelson,1995). We believe that the more general "agnostic" learning scenario and the noisy learning problem are related. It seems useful, in this context, to extend the size of the committee and use more refined definitions for "disagreement" among the committee members.

In this work we have explored some of the power of algorithms for learning using queries that have access to random unlabeled instances and can make membership queries. This model of learning is natural in contexts where unlabeled instances are much cheaper than labeled instances. An interesting theoretical open question is how much more powerful is this model of learning from queries from the standard model for using membership queries in statistical learning.

Acknowledgments

Part of this research was done at the Hebrew University of Jerusalem. Freund, Shamir and Tishby would like to thank the US-Israel Binational Science Foundation (BSF) Grant no. 90-00189/2 for support of their work. We would also like to thank Yossi Azar, Shlomo Halfin, and Manfred Opper for helpful discussions regarding this work.

Table 1. Notation Table

symbol	definition	meaning	section	equation
X		sample space	2	
$\mathcal{D}$		sample distribution	2	
$\vec{X}$	$\{x_1, x_2, \ldots\}$	unlabeled examples drawn from X according to $\mathcal{D}$	2 2	
m		number of examples	2	
n		number of queries (labeled examples)	2	
$\vec{X}_{1\ldots m}$	$\{x_1, \ldots, x_m\}$	first m examples	2	
$\mathbf{C}$		concept class	2	
c		target concept	2	
V_n	$\{h \in C \mid h(x_i) = c(x_i), i = 1 \ldots n\}$	version space of first n labeled examples	2	
$\mathcal{P}$		Bayesian prior distribution on $\mathbf{C}$	2	
h		hypothesis in $\mathbf{C}$	2	
$\mathcal{I}$	$-\log \Pr_{\mathcal{P}}(V_m)$	cumulative information gain	2	1
$\mathcal{H}(p)$	$-p \log p - (1-p) \log(1-p)$	binary entropy function	2	
$\mathcal{G}(x_i \mid V_{i-1})$		expected information gain from example x_i given version space V_{i-1}	2	2
F		fractional reduction in version space	4	
M	$\{1, \ldots, m\}$		5	
$\vec{X}_M$	$\{x_1, \ldots, x_m\}$	first m examples in X	5	
I	$\{i_1, i_2, \ldots\}$	sequence of indices of examples used as queries	5	
I_n	$\{i_1, \ldots, i_n\}$	first n elements of I	5	
$\vec{X}_I$	$\{x_{i_1}, x_{i_2}, \ldots\}$	sequence of query examples	5	
$\vec{X}_{I_n}$	$\{x_{i_1}, \ldots, x_{i_n}\}$	first n examples used as queries	5	
g		lower bound on expected information gain	5	
d		VC dimension	5	
G		expected information gain functional	6.1	18
K	$F(1-F)$		6.1	
$Q(x)$	$\mathcal{H}(1/2 - \sqrt{(1-4x)/2})$		6.1	
Ψ		variation in K	6.1	
$\lambda_{\mathcal{P}}, \lambda_{\mathcal{D}}$		uniformity parameters of prior and input distributions	6.2	
λ_c		divergence between correct and incorrect priors	6.3	

Appendix A

Proofs of Lemmas 4-7

Proof of Lemma 4: Let us denote by S_t the convex body in R^{d-1} that is defined by the slice of the convex body B at t. Clearly, $f(t)$ is the volume of S_t.

We define the linear combination of two bodies, A and B as:

$$\lambda_1 A + \lambda_2 B = \{\lambda_1 a + \lambda_2 b | a \in A,\ b \in B\} ,$$

where $\lambda_1, \lambda_2 \in R$. An immediate result of the convexity of B is that for any $t_1, t_2 \in R$, and any $0 \leq \lambda_1, \lambda_2 \leq 1$ such that $\lambda_1 + \lambda_2 = 1$

$$\lambda_1 S_{t_1} + \lambda_2 S_{t_2} \subseteq S_{\lambda_1 t_1 + \lambda_2 t_2} .$$

Using the terminology of the theory of convex bodies, we can say that the set of bodies S_t, parameterized by $t \in R$ is a (one-parameter) concave family of bodies.[12]

The Brunn-Minkowski theorem states that, for bodies in R^n, *"the n-th root of the volume of the bodies of a linear or concave family is a concave function of the family of parameters"* ((Bonnesen & Fenchel,1987),Subsection 48). In our case, $n = d - 1$ and the family is a concave family of a single parameter. We thus get the statement of the lemma as a special case of the Brunn Minkowski theorem.

■

Proof of Lemma 6: As the value of the functional $G(F)$ is always positive, there must exist an infimum to the set of values it can achieve on the set of all convex bodies. We denote this infimum by μ and show that it is achieved as a minimum. In other words, that there exists a volume function F_∞ which corresponds to a convex body such that $G(F_\infty) = \mu$.

Let B_n be a sequence of convex bodies and F_n be the corresponding sequence of volume functions such that $\lim_{n\to\infty} G(F_n) = \mu$. By Lemma 4, we may assume that the bodies B_n are bodies of revolution, and that they correspond to concave radius functions $r_n(t)$. We thus need to show that there exists a concave radius function $r_\infty(t)$ which is the limit of $r_n(t)$ for $n \to \infty$.

The functional $G(F)$ is defined in terms of integrals and the radius functions $r_n(t)$ are continuous and bounded by a constant which depends only on the dimension d. Thus if $r_n(t)$ converges to $r_\infty(t)$ pointwise then the value of G on the sequence of bodies of revolution corresponding to $r_n(t)$ converges to the value of G on the body corresponding to $r_\infty(t)$.

We prove the lemma by showing the existance of a subsequence of the radius functions which have a pointwise limit. Using a diagonalization argument, we can pick a subsequence of r_n, indexed by m, such that $r_m(t)$ converges pointwise for each rational value of t. It is easy to see that the limit function $r_\infty(t)$, defined on the rationals, is concave and continuous there. We get a concave extension to all real values of t by taking the limit over the rationals:

$$r_\infty(t_0) = lub(r(\tau)|\tau < t_0, \tau \text{ rational }) \ .$$

Clearly $r_\infty(t)$ is also concave and continuous and is the pointwise limit of $r_n(t)$ for all t. Thus $r_\infty(t)$ is the radius function of a concave body B which assumes the minimum $G(B) = \mu$. ■

Proof of Lemma 7: The radius function that corresponds to the two-cone body is

$$r^*(t) = c_d \max(0, 1 - |t|) \tag{A.1}$$

One can compute $G_d(r^*)$ for any fixed d by solving the integral in Equation (18) as follows. In this case we find it more convenient to use the integral over the negative half of the line as defined in Equation (19). The volume function in the range $-1 \le t \le 0$ is $F_d^*(t) = \int_{-\infty}^{t} (r^*(s))^{d-1} ds = (1+t)^d/2$ and it is 0 for $t < 0$. Plugging this into Equation (19) we get

$$G(F_d^*) = \frac{\int_0^1 \frac{(1+t)d+1}{2}(1 - \frac{(1+t)d+1}{2})\mathcal{H}(\frac{(1+t)d+1}{2})\, dt}{\int_0^1 \frac{(1+t)d+1}{2}(1 - \frac{(1+t)d+1}{2})\, dt} = \frac{\int_0^{1/2} F^{1/d}(1-F)\mathcal{H}(F)dF}{\int_0^{1/2} F^{1/d}(1-F)dF} \ , \tag{A.2}$$

which can be shown by direct calculation to decrease as $d \to \infty$. Which gives the general lower bound of

$$G(F_d^*) > \frac{\int_0^{1/2}(1-F)\mathcal{H}(F)dF}{\int_0^{1/2}(1-F)dF} = \frac{1}{9} + \frac{7}{18\log 2} \ . \tag{A.3}$$

This proves the statement of the lemma. ■

Proof of Lemma 5: We shall keep using the notation defined in the proof of Theorem 2. For each volume function F which does not come from a cone, we construct a variation that decreases $G(F)$.

We describe the variations in terms of adding a variation function, $\psi(t)$ to the radius function $r(t)$. As we are restricting ourselves to volume functions, it is enough to define $\psi(t)$ for $0 \le t < \infty$.

Let us enumerate the requirements on the radius variation function $\psi(t)$, and on the corresponding volume variation function
$F(t) + \Psi(t) = c_{d-1} \int_0^t (r(s) + \psi(s))^{d-1}\, ds$.

1. We need $F(|t|) + \Psi(|t|)$ to be a volume function. For this to hold we require that $r(t) + \psi(t)$ is a positive concave function that is nonzero only on a bounded segment $[0, c]$, $c < \infty$.

2. We need to guarantee that $\int_0^\infty \nabla G(t)\Psi(t)\,dt < 0$. For that to hold we require that $\Psi(t)$ is non-positive for all $0 \leq t \leq w$ and non-negative for all $t > w$. Where w is the pivot point for the volume function F. See equation 20 and the discussion following it.

3. For any given $\epsilon > 0$ we should be able to find a radius variation function $\psi(t)$ such that the change in the corresponding volume function is as small as is desired $\epsilon > \int_0^{+\infty} \Psi(t)^2\,dt > 0$.

We describe three families of variational functions. For any radius function r that corresponds to a volume function and is not equal to $r^* = \max(0, 1-|t|)$, one of these variations applies, showing that there exists $r'(t)$ such that $G_d(r') < G_d(r)$. The variations are constructed geometrically. Below is a list of the constructions that should be read alongside Figure A.1. The basic idea in all three transformations is to "move" volume from place to place along the projection direction, in such a way that for each point t in a particular range, volume is moved only from one the right of the points to their left or vice versa. It is easy to check that each of the conditions 1-3 holds for each of those transformations. In the descriptions below we shall refer to volume changes are caused by increasing or decreasing the radius function, note that these are changes in the d-dimensional volume of the revolution body whose volume function corresponds to the radius function, and not in the two dimensional area described by the changes in the graph. The transformations thus depend on the dimension of the actual body, however, the qualitative form of the transformation remains the same for all dimensions. Each transformation takes a parameter λ, which is a positive number that is set small enough so that condition 3 holds.

1. If r is not linear in the range $0 \leq t \leq w$ then transformation 1 is used (see Figure A.1(a)):

   (A) Let A be the point $(w, r(w))$, select a point A' on the curve defined by r to the left of A so that the volume decrease caused by changing the curve[13] $A \frown A'$ to the chord $A - A'$ is equal $\lambda/2$.

   (B) Let B be the point $(0, r(0))$, select a point B' slightly above B and connect it to the (unique) point X on the curve so that the curve $B - X \frown A' - A$ is concave. Choose B' so that the volume increase caused by changing the curve $B \frown X$ to the line $B - X$ is $\lambda/2$.

   Set λ_0 small enough so that this construction is possible for all $0 < \lambda < \lambda_0$.

   Note that for each point $0 \leq t \leq w$, at least one of the two following conditions hold: either volume is only removed from the right of t, or volume is only added to the left of t. This implies that the volume function, $F(t)$, increases in this range. Because the amount of volumes that are removed and added are equal, $F(t)$ does not change for t outside the range $[0, w]$. This implies that condition 2 holds.

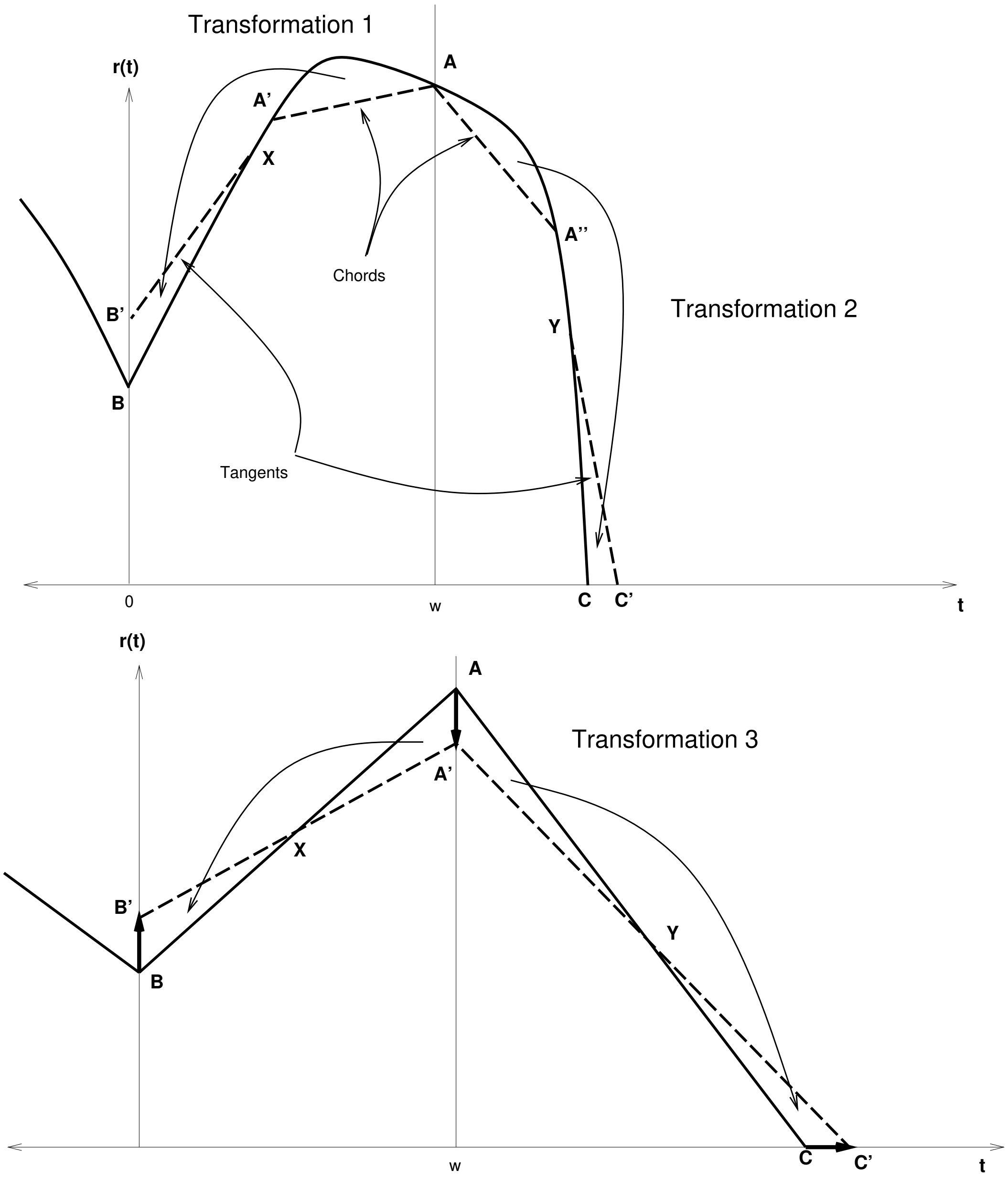

Figure A.1. The variational transformations

2. If r does not decrease linearly to zero for $t \geq w$ then transformation 2 is used (see Figure A.1(a)):

   (A) Select A'' on the curve to the right of A so the volume decrease that is caused by changing $A \frown A''$ to $A - A''$ is $\lambda/2$.

   (B) Let C be the point at which the curve meets the horizontal axis. Select C' slightly to the right of C and connect it to the point Y on the curve so that the curve $C' - Y \frown A'' - A$ is concave. Choose C' so that the volume increase caused by changing $C' - C - Y$ to $C' - Y$ is $\lambda/2$.

   Set λ_0 small enough so that this construction is possible for all $0 < \lambda < \lambda_0$.

   An argument similar to the one used in transformation 1 holds in this case for $t > w$.

3. If neither condition 1 nor 2 holds, and the slopes of the two linear segments are not equal (i.e. $r \neq r^*$), then transformation 3 is used (see Figure A.1(b)):

   (A) A point A' slightly below A is chosen.

   (B) A point B' slightly above B is chosen so that there is no net change in the volume when changing $A - B$ to $A' - B'$.

   (C) A point C' slightly to the right of C is chosen so that there is no net change in the volume when changing $A - C$ to $A' - C'$.

   (D) The movement from A to A' is chosen do that the change in the volume caused by each of the four changes in r: $B - X$ to $B' - X$, $A - X$ to $A - X'$, $A - Y$ to $A' - Y$ and $C - Y$ to $C' - Y$ is equal to $\lambda/4$

   In this case the volume function is changed on both sides of the pivot point. Arguments similar to the one used in transformation 1 shows that condition 2 is met.

The only radius functions to which none of those transformations apply is r^*, thus finishing the proof of the lemma. ∎

Appendix B

Proof of Theorem 3

We first prove the dependence on the uniformity of the input distribution, as measured by $\lambda_{\mathcal{D}}$. In general, any distribution $\mathcal{D}$ that is within $\lambda_{\mathcal{D}}$ of the uniform distribution μ can be written as weighted sum of the form $\lambda_{\mathcal{D}}\mu + (1 - \lambda_{\mathcal{D}})\nu$ where ν is some other distribution. Fix the version space and *any* prior distribution, let the distribution of examples be $\mathcal{D} = \lambda_{\mathcal{D}}\mu + (1 - \lambda_{\mathcal{D}})\nu$ and let g_μ, g_ν be the expected information gains when the examples are generated according to μ or ν respectively. As $g_\nu > 0$ we get that the expected information gain when $\mathcal{D}$ is within $\lambda_{\mathcal{D}}$ of uniform is at least $\lambda_{\mathcal{D}}$ times the expected information gain when $\mathcal{D}$ is uniform.

The analysis of the dependence on $\lambda_{\mathcal{P}}$ is more involved. We go back to the analysis of an arbitrary projection of a convex body from the proof of Theorem 2. The main idea there was to show transformations that increase or decrease the volume function in particular ranges, in a way that decreased the expected information gain. There, the transformation involved changing the shape of the body. Here we present a transformation that changes the density of the prior distribution inside the version space.

We fix a convex body and a direction $\vec{x}$ along which this body is projected. We denote by $\rho(t)$ the average density along the slice of body which is defined by the example $(\vec{x}, t)$. The relation between the volume function F, and the radius function r is now

$$F_d(t) = \int_{-\infty}^{t} (r(s))^{d-1} \rho(s) ds \ .$$

We search for a density distribution of the points in the body, which is within $\lambda_{\mathcal{P}}$ of the uniform distribution, and minimizes the expected information gain from (uniformly distributed) examples whose direction is $\vec{x}$. Note that the symmetrization argument used in the proof of Theorem 2 holds for this case too, and we can thus restrict ourselves to functions r and ρ that are defined only over the positive reals. From the variational derivative of $F(t)$ for $t \geq 0$ that we computed in Equation (20), we know that $G(F)$ decreases if $F(t)$ is increased for some $t \leq w$ or if $F(t)$ is decreased for some $0 \leq t \leq w$. As we allow deviations from the uniform prior distribution we can change F without changing the form of the convex body. We shall now give a variation of ρ that changes $\rho(t)$ in the range $0 \leq t \leq w$ in a way that decreases $G(F)$. As this variation can be applied to any ρ that does not have a specific step-like form in this range, we get that this step-like form of ρ achieves the minimal value of $G(F)$ for this fixed body and $\mathcal{P}$ that is within $\lambda_{\mathcal{P}}$ of uniform. A similar argument can be used to show that $\rho(t)$ must also have a stepwise form in the range $w \leq t$.

Assume that there exist $0 < t_1 < t_2 < w$ and $\epsilon, \delta > 0$ such that $0 \leq t_1 - \epsilon < t_1 + \epsilon \leq t_2 - \epsilon < t_2 + \epsilon < w$, and such that for all $t \in [t_1 - \epsilon, t_1 + \epsilon]$, $\rho(t) < 1/\lambda_{\mathcal{P}} - \delta$, and for all $t \in [t_2 - \epsilon, t_2 + \epsilon]$, $\rho(t) > \lambda_{\mathcal{P}} + \delta$. We add to $\rho(t)$ the following variation function:

$$\psi(t) = \begin{cases} +\delta_1, & t_1 - \epsilon \leq t \leq t_1 + \epsilon \ , \\ -\delta_2, & t_2 - \epsilon \leq t \leq t_2 + \epsilon \ , \\ 0, & \text{otherwise} \end{cases} \ ,$$

where δ_1, δ_2 are chosen so that $\delta \geq \delta_1, \delta_2 > 0$ and

$$\frac{\delta_1}{\delta_2} = \frac{\int_{t_1-\epsilon}^{t_1+\epsilon} (r(s))^{d-1} ds}{\int_{t_2-\epsilon}^{t_2+\epsilon} (r(s))^{d-1} ds} \ .$$

This insures that the volume function does not change outside the range $[t_1-\epsilon, t_2+\epsilon]$.

It is easy to check that $\rho(t)+\psi(t)$ corresponds to a density distribution that is within $\lambda_{\mathcal{P}}$ of the uniform distribution. Changing the density distribution from $\rho(t)$ to $\rho(t)+\psi(t)$ decreases $F(t)$ in the range $[t_1-\epsilon, t_2+\epsilon]$ and does not change $F(t)$ anywhere else. Thus this change decreases $G(F)$. It is also easy to check that this variation cannot be applied to ρ if and only if there exists $0 \leq a \leq w$ such that $\rho(t) = 1/\lambda_{\mathcal{P}}$ for $0 \leq t < a$ and $\rho(t) = \lambda_{\mathcal{P}}$ for $a < t \leq w$. From this argument and a similar argument for the range $t \geq w$ we get that the density function that minimizes $G(F)$ must be of the form

$$\rho^*(t) = \begin{cases} 1/\lambda_{\mathcal{P}}, & 0 \leq t \leq a \text{ or } b \leq t\ , \\ \lambda_{\mathcal{P}}, & a \leq t \leq b \end{cases} . \tag{B.1}$$

where $0 \leq a \leq w \leq b$. We do not have a simple variational argument for determining the exact value of a and b, however, as we shall see, we can lower bound the information gain without this explicit knowledge.

We have thus found the form of the density function that minimizes the information gain for a specific body (and a specific projections). Suppose now that we fix the function ρ and vary the shape of the body, i.e. the radius function r. Going through the construction of the variational functions ψ in the proof of Theorem 2, we see that the same construction steps hold verbatim, although special attention needs to meaning of the expression "the volume decrease is equal to x" as the volume is now defined in terms of the non uniform distribution specified by ρ.

The combination of these two arguments shows that the smallest value of $G(F)$ is attained for the radius function r^* specified in Equation (A.1), and the average density function ρ^*. It remains to compute a lower bound on $G(F)$ based on these two facts. This is done by bounding the ratio between the values of $G(F)$ for the uniform prior and the non uniform prior cases.

We change the integration variable in Equation (19) from x to $F(x)$:

$$G(F) = \frac{\int_0^{1/2} F(1-F)\, \mathcal{H}(F)\, \frac{dx}{dF} dF}{\int_0^{1/2} F(1-F)\, \frac{dx}{dF} dF} . \tag{B.2}$$

When written in this form, the dependence of $G(F)$ on the r and ρ enters the equation through the derivative dx/dF. By bounding the ratio between the values that this derivative attains in the uniform and the non-uniform cases, we can bound the ratio between the values that $G(F)$ attains for the uniform and the non-uniform prior distributions.

The volume function that corresponds to the uniform prior distribution is, for $-1 \leq x \leq 0$, $F_{\text{unif}}(x) = (1+x)^d/2$. The volume function that corresponds to the prior distribution defined by ρ^* is

$$F_{\text{non-unif}}(x) = \frac{1}{2} \begin{cases} \lambda_{\mathcal{P}}^{-1}(1+x)^d, & -1 \leq x \leq -b, \\ \lambda_{\mathcal{P}}(1+x)^d + c, & -b \leq x \leq -a, \\ \lambda_{\mathcal{P}}^{-1}(1+x)^d + 1 - \lambda_{\mathcal{P}}^{-1}, & -a \leq x \leq 0 \end{cases} \tag{B.3}$$

Where $c \geq 0$ is defined by matching the two definitions of $F(-b)$.

Taking the derivatives of F_{unif} and $F_{\mathrm{non\text{-}unif}}$ we get the following equation for their ratio:

$$\frac{\left(\frac{dx}{dF}\right)_{\mathrm{non\text{-}unif}}}{\left(\frac{dx}{dF}\right)_{\mathrm{unif}}} = \begin{cases} \lambda_{\mathcal{P}}^{1/d}, & 0 \leq F \leq F(-b), \\ \lambda_{\mathcal{P}}^{-1/d}\left(\frac{2F}{2F-c}\right)^{1-1/d}, & F(-b) \leq F \leq F(-a), \\ \lambda_{\mathcal{P}}^{1/d}\left(\frac{2F}{2F+\lambda_{\mathcal{P}}^{-1}-1}\right)^{1-1/d}, & F(-a) \leq F \leq 1/2 \end{cases} \tag{B.4}$$

Using the facts that $\lambda_{\mathcal{P}} \leq 1$, $c \geq 0$, and $d \geq 2$ we can bound the ratio of the derivatives for each of the three cases. For the range $-1 \leq x \leq -b$ we get that

$$\lambda_{\mathcal{P}} \leq \lambda_{\mathcal{P}}^{1/d} \leq \frac{\left(\frac{dx}{dF}\right)_{\mathrm{non\text{-}unif}}}{\left(\frac{dx}{dF}\right)_{\mathrm{unif}}} \leq 1 \ . \tag{B.5}$$

For the range $-b \leq x \leq -a$ we get, using the fact that F is monotone non-decreasing, that

$$1 \leq \frac{2F(-a)}{2F(-a)-c} \leq \frac{2F(x)}{2F(x)-c} \leq \frac{2F(-b)}{2F(-b)-c} = \frac{\lambda_{\mathcal{P}}^{-1}(1-b)^d}{\lambda_{\mathcal{P}}(1-b)^d} \leq \lambda_{\mathcal{P}}^{-2} \ ,$$

which implies that in the range $-b \leq x \leq -a$,

$$1 \leq \lambda^{-1/d} \leq \frac{\left(\frac{dx}{dF}\right)_{\mathrm{non\text{-}unif}}}{\left(\frac{dx}{dF}\right)_{\mathrm{unif}}} \leq \lambda_{\mathcal{P}}^{-2+1/d} \leq \lambda_{\mathcal{P}}^{-2} \ . \tag{B.6}$$

Finally, for the range $-a \leq x \leq 0$, we get that

$$\lambda_{\mathcal{P}}^2 \leq \frac{\lambda_{\mathcal{P}}(1-a)^d + c}{\lambda_{\mathcal{P}}^{-1}(1-a)^d} = \frac{2F(-a)}{2F(-a)+\lambda_{\mathcal{P}}^{-1}-1} \leq 1$$

which implies that

$$\lambda_{\mathcal{P}}^2 \leq \lambda_{\mathcal{P}}^{2-1/d} \leq \lambda^{-1/d} \leq \frac{\left(\frac{dx}{dF}\right)_{\mathrm{non\text{-}unif}}}{\left(\frac{dx}{dF}\right)_{\mathrm{unif}}} \leq 1 \tag{B.7}$$

Combining the bounds from Equations (B.5), (B.6), and (B.7), and plugging them into Equation (B.4), we get that

$$\lambda_{\mathcal{P}}^2 \leq \frac{\left(\frac{dx}{dF}\right)_{\mathrm{non\text{-}unif}}}{\left(\frac{dx}{dF}\right)_{\mathrm{unif}}} \leq \lambda_{\mathcal{P}}^{-2}$$

Using this bound and Equation (B.2) we get that $G(F_{\mathrm{non\text{-}unif}}) \geq \lambda_{\mathcal{P}}^4 G(F_{\mathrm{unif}})$. This completes the proof of the theorem.

Notes

1. Here, and elsewhere in the paper, $\log(\cdot)$, denotes the logarithm over base two, while $\ln(\cdot)$ denotes the logarithm over base e.
2. Our analysis can be extended to larger committees, but the improvement in the performance is minor.
3. For example, consider the case in which the version space contains two disconnected sets in R^2, which are very far from each other, and assume that a random example is very likely to separate these two sets. Suppose one of the sets has probability ϵ, while the other has probability $1 - \epsilon$. While most of the examples that separate the two sets are rejected, the fraction that is accepted can still dominate all other examples. Thus the expected information gain is close to $\mathcal{H}(\epsilon)$. As ϵ can be set arbitrarily small, the expected information gain can be arbitrarily close to zero. It seems that this type of version space can occur only very rarely but we do not know what are the necessary conditions.
4. Note that the number of calls to **Sample** is $\Omega(d/\epsilon)$ ((Blumer et al.,1989)), even if *all* of the instances are used as queries to **Label**.
5. The bound as it appears in (McDiarmid,1989) is given for martingales. However, it is easily checked that it is also true for super-martingales. Reversing the sign of the Y_i we get an equivalent theorem for sub-martingales.
6. Actually, it is enough to assume that the distribution of the offset t is uniform for any direction $\vec{x}$. No assumption needs to be made regarding the distribution of $\vec{x}$.
7. The perceptron concept class is defined as the following set of binary functions over the unit sphere

$$c_{\vec{w},t}(\vec{x}) \begin{cases} 1, & \vec{x}\cdot\vec{w} \geq t \\ 0, & \text{otherwise} \end{cases} .$$

8. In the following discussion we ignore the distinction between the concepts in **C** and their parameterization, and refer to the concept $c_{\vec{w}}$ simply as the vector $\vec{w}$.
9. Details on how the Fréchet derivative is defined and calculated can be found in standard books on variational analysis, such as (Smith,1985).
10. This definition is a one-sided version of the notion of λ-closeness defined in Definition 6.2.
11. Ignoring log factors.
12. For the definition of a convex family of bodies see ((Bonnesen & Fenchel,1987),Subsection 24).
13. We use $A - B$ to denote the line segment between the points A and B, and $A \frown B$ to denote the segment of a curve that connects A and B. We also use the shorthand $A - B \frown C - D$ to denote a the concatenation of a line segment, a curve segment, and another line segment.

References

Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, April 1988.

A. C. Atkinson and A. N. Donev. *Optimum Experimental Designs*. Oxford science publications, 1992.

Ian Barland. Some ideas on learning with directional feedback. Master's thesis, University of California at Santa Cruz, June 1992.

E. Baum. Neural net algorithms that learn in polynomial time from examples and queries. *IEEE Transactions on Neural Networks*, 2:5–19, 1991.

E. B. Baum and K. Lang. Query learning can work poorly when a human oracle is used. In *International Joint Conference in Neural Networks*, Beijing, China, 1992.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

T. Bonnesen and W. Fenchel. *Theory of Convex Bodies.* BCS Associates, Moscow, Idaho, USA, 1987.

David Cohn, Les Atlas, and Richard Ladner. Training connectionist networks with queries and selective sampling. In D. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, San Mateo, CA, 1990. Morgan Kaufmann.

Ido Dagan and Sean P. Engelson. Committee-based sampling for training probabilistic classfiers. In Priedits and Russel, editors, *The XII International Conference on Machine Learning*, pages 150–157. Morgan Kaufmann, 1995.

Bonnie Eisenberg and Ronald L. Rivest. On the sample complexity of pac-learning using random and chosen examples. In *Proceedings of the 1990 Workshop on Computational Learning Theory*, pages 154–162, 1990.

V. V. Fedorov. *Theory of Optimal Experiments.* Academic Press, New York, 1972.

David Haussler, Michael Kearns, and Robert E. Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning*, 14:83–113, 1994.

W. Kinzel and P. Ruján. Improving a network generalization ability by selecting examples. *Europhys. Lett.*, 13:473–477, 1990.

D. V. Lindley. On a measure of the information provided by an experiment. *Ann. Math. Statist.*, 27:986–1005, 1956.

C. McDiarmid. On the method of bounded differences. In *Survey of Combinatorics, 10th British Combinatorial Conference*, 1989.

T.M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2), 1982.

N. Sauer. On the density of families of sets. *J. Combinatorial Theory (A)*, 13:145–147, 1972.

H.S Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Workshop on Computational Learning Theory*, pages 287–294, San Mateo, CA, 1992. Morgan Kaufmann.

Peter Smith. *Convexity Methods in Variational Calculus.* Research studies press, John Wiley & sons, 1985.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

Received Date
Accepted Date
Final Manuscript Date