Theory of Disagreement-Based Active Learning

Steve Hanneke
steve.hanneke@gmail.com

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

S. Hanneke. *Theory of Disagreement-Based Active Learning*. Foundations and Trends® in Machine Learning, vol. 7, no. 2-3, pp. 131–309, 2014.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-809-6

Foundations and Trends® in Machine Learning
Volume 7, Issue 2-3, 2014

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles on the theory, algorithms and applications of machine learning, including the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive, and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2014, Volume 7, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in Machine Learning
Vol. 7, No. 2-3 (2014) 131–309

DOI: 10.1561/2200000037

now
the essence of knowledge

Theory of Disagreement-Based Active Learning

Steve Hanneke
steve.hanneke@gmail.com

Contents

Abstract

Active learning is a protocol for supervised machine learning, in which a learning algorithm sequentially requests the labels of selected data points from a large pool of unlabeled data. This contrasts with passive learning, where the labeled data are taken at random. The objective in active learning is to produce a highly-accurate classifier, ideally using fewer labels than the number of random labeled data sufficient for passive learning to achieve the same. This article describes recent advances in our understanding of the theoretical benefits of active learning, and implications for the design of effective active learning algorithms. Much of the article focuses on a particular technique, namely disagreement-based active learning, which by now has amassed a mature and coherent literature. It also briefly surveys several alternative approaches from the literature. The emphasis is on theorems regarding the performance of a few general algorithms, including rigorous proofs where appropriate. However, the presentation is intended to be pedagogical, focusing on results that illustrate fundamental ideas, rather than obtaining the strongest or most general known theorems. The intended audience includes researchers and advanced graduate students in machine learning and statistics, interested in gaining a deeper understanding of the recent and ongoing developments in the theory of active learning.

S. Hanneke. *Theory of Disagreement-Based Active Learning.* Foundations and Trends® in Machine Learning, vol. 7, no. 2-3, pp. 131–309, 2014.

DOI: 10.1561/2200000037.

1

Introduction

Active learning is a general protocol for supervised machine learning, involving interaction with an expert or oracle. Though there are many variants of active learning in the literature, the focus of this article is the so-called *pool-based* active learning model. Specifically, we suppose the user has obtained a (typically large) number of unlabeled data points (i.e., only the features, or covariates, are present), referred to as the unlabeled *pool*. The learning algorithm is permitted complete access to these unlabeled data. It additionally has access to an expert or oracle, capable of providing a label for any instance in this pool upon request, where the label corresponds to the concept to be learned. The queries to this expert can be sequential, in the sense that the algorithm can observe the responses (labels) to its previous requests before selecting the next instance in the pool to be labeled. As is typically the case in supervised machine learning, the objective is to produce a classifier such that, if presented with fresh unlabeled data points from the same data source, the classifier would typically agree with the label the expert would produce if he or she were (hypothetically) asked. We are especially interested in algorithms that can achieve this objective without requesting too many labels from the expert. In this regard,

the active learning protocol enables us to design more powerful learning methods compared to the traditional model of supervised learning (including semi-supervised learning), here referred to as *passive learning*, in which the data points to be labeled by the expert are effectively selected at random from the pool. Indeed, the driving question in the study of active learning is how many fewer labels are sufficient for an active learning algorithm to achieve a given accuracy, compared to the number of labels necessary for a passive learning algorithm to achieve the same.

The motivation for active learning is that, in many machine learning problems, unlabeled data are quite inexpensive to obtain in abundance, while labels require a more time-consuming or resource-intensive effort to obtain. For instance, consider the problem of webpage classification: say, classifying a webpage as being about "news" or not. A basic web crawler can very quickly collect millions of web pages, which can serve as the unlabeled pool for this learning problem. In contrast, obtaining labels typically requires a human to read the text on these pages to determine whether it is a news article or not. Thus, the time-bottleneck in the data-gathering process is the time spent by the human labeler. It is therefore desirable to minimize the number of labels required to obtain an accurate classifier. Active learning is a natural approach to doing so, since we might hope to reduce the amount of redundancy in the labels provided by the expert by only asking for labels that we expect to be, in some sense, quite informative, given the labels already provided up to that time.

1.1 Why Do We Need a Theory of Active Learning?

The potential for active learning to achieve accuracies comparable to passive learning using fewer labels has been observed in many practical applications over the past several decades. However, intermixed with these shining positive outcomes has been an equally-vast array of applications for which these same active learning methods failed to provide any benefits; some of these algorithms have even been observed to perform *worse* than their passive learning counterparts in certain appli-

cation domains. How should we interpret these negative outcomes? Is the active learning protocol fundamentally unable to provide any benefits in these application domains, or might these observations simply reflect the need to develop smarter active learning algorithms? Questions such as these beg for a theoretical treatment. More abstractly, we are asking what kind of performance we should expect from a well-designed active learning algorithm, so that we may evaluate whether a given method meets this standard. Is it reasonable to expect an algorithm to always provide improvements over passive learning, or will there be some applications where no active learning strategy can outperform a given passive learning strategy? In the scenarios where active learning is potentially beneficial, how many fewer labels should we expect a well-designed active learning algorithm to require for obtaining a given accuracy? Attempts to answer these questions naturally lead us to a deeper understanding of the general principles that should underly well-designed active learning algorithms, so that the result of such an investigation is both a better understanding of the fundamental capabilities of active learning, and insights that can guide the design of practical active learning algorithms.

A second motivation for developing a theory of active learning is that, as will hopefully be apparent in the presentation below, many wonderfully beautiful and elegant mathematical concepts and theorems arise quite naturally out of the active learning formalism. We are incredibly lucky that such a natural framework for interactive machine learning can be studied in such generality, with many general properties concisely characterized by such simple mathematical constructions. For reasons such as these, the exploration of this fascinating mathematical landscape has become a source of satisfaction and joy for many in the growing community of active learning researchers.

1.2 What is Covered in This Article?

This article includes some of the recent advances in the theory of active learning, focusing on characterizing the number of label requests sufficient for an active learning algorithm to achieve a given accuracy;

this number is known as the *label complexity*. As our interest in active learning is in its ability to reduce the label complexity compared to passive learning, we will also review some of the known results for passive learning, so as to establish a baseline for comparison.

Throughout much of the article, we will focus on one particular active learning technique, known as *disagreement-based* active learning. The reason for this choice is that the literature on disagreement-based active learning represents a fairly coherent, elegant, and mature thread in the broader active learning literature, and is now quite well-understood, with a rich variety of established results. It provides us a unified approach to active learning, which can be applied with essentially any classifier representation, can be studied under a variety of noise models, and composes well with standard relaxations that enable computational efficiency (namely, the use of surrogate losses). The established results bounding the label complexity of this technique are concise, easy to comprehend, and often fairly tight (in the sense that the algorithm actually requires nearly that many labels).

However, it is known that disagreement-based active learning is sometimes not optimal. For this reason, we additionally discuss several alternative techniques, most of which are more involved and less understood, but which are known to sometimes yield smaller label complexities than disagreement-based methods. As the literature on these other techniques is less developed, our discussion of each of them will necessarily be somewhat brief; however, some of these approaches represent important directions for investigation, and further development of these techniques would undoubtedly be of great value.

The basic outline of the article is as follows. Chapter 2 introduces the formal setting, some basic notation, and essential definitions, along with a few basic examples illustrating the fundamental concepts, style of analysis, and typical results. Chapter 3 briefly surveys the known results on the label complexity of passive learning, which serve as a baseline for comparison throughout. Chapter 4 describes several known lower bounds on the label complexity of active learning, which provide an additional point of comparison, particularly in discussions of optimality. Chapter 5 introduces the basic idea of disagreement-

based active learning, along with a thorough analysis of the technique for the simple scenario of noise-free learning (the so-called *realizable case*). This is followed by a description of a noise-robust variant of the disagreement-based learning strategy, and an analysis of its label complexity under various commonly-studied noise conditions. In Chapter 6, we discuss a simple trick, involving the use of a convex relaxation of the loss function, which can make the previously-discussed algorithm computationally efficient, while still allowing us to provide formal guarantees on its label complexity under certain restricted conditions. The results concerning the label complexity of disagreement-based active learning are expressed in terms of a simple quantity, known as the *disagreement coefficient*. Chapter 7 is dedicated to describing the known properties of the disagreement coefficient, including sufficient conditions for it to obtain favorable values, and several specific learning problems for which the value of the disagreement coefficient has been calculated. Finally, Chapter 8 briefly surveys several of the other threads from the literature on the theory of active learning. It is worth mentioning that the dependences among several of these chapters are rather weak. In particular, most of the discussion of bounds on the disagreement coefficient in Chapter 7 can be read anytime after Chapter 2. Additionally, the discussion of surrogate losses in Chapter 6 can be considered largely optional in the sequence, and may be skipped without significant loss of continuity (aside from dependences in Section 8.8).

Much of the article is structured around a few algorithms, emphasizing several theorems concerning their respective label complexities, along with a variety of results on the relevant quantities those results are expressed in terms of. Where appropriate, I have accompanied these results with rigorous proofs. However, as this discussion is intended to be pedagogical, in many cases I have refrained from presenting the strongest or most general form of the results from the literature, instead choosing a form that clearly illustrates the fundamental ideas without requiring too many additional complications; the article includes numerous references to the literature where the interested reader can find the stronger or more general forms of the results. I have also attempted to provide high-level reasoning for each of the main results, so that ca-

sual readers can grasp the core ideas motivating the algorithms and leading to the formal theorems, without needing to wade through the details needed to convert the ideas into a formal proof. The technical content of this article is intended to be suitable for researchers and advanced graduate students in statistics or machine learning, familiar with the basics of probability theory and statistical learning theory at the level of an introductory graduate course.

Remark The present article is an abbreviated version of a longer manuscript [Hanneke, 2014], which can be downloaded from the author's website. Some of the additional material in the extended version is referenced in the chapters below. Additionally, the long version may be updated from time to time as this field continues to develop.

1.3 Conceptual Themes

Before beginning the technical discussion, we first briefly illustrate some of the main concepts that arise below. Readers completely unfamiliar with active learning may also find the brief survey of Dasgupta [2011] helpful, as it provides a concise and lucid description of the main themes, without getting into as much technical detail as the present article.

As mentioned, the focus of much of this article is on the strategy of *disagreement-based* active learning, an elegant and general idea introduced in the seminal work of Cohn, Atlas, and Ladner [1994]. To illustrate this idea, consider the problem of learning a *linear separator* in the 2-dimensional plane: that is, the label of each point is "+" if the point is on one side of a particular (unknown) line, called the *target separator*, and is "−" if the point is on the other side. Suppose, at some time, we have observed a few labeled data points, as in Figure 1.1a. We know the target separator is some line that separates all of the "+" points from the "−" points; a few such lines are depicted in Figure 1.1b (in truth, there are an infinite number of possibilities). If we are then given a new unlabeled point, such as the one marked "○" in Figure 1.1c, the question is whether or not we should request its label.

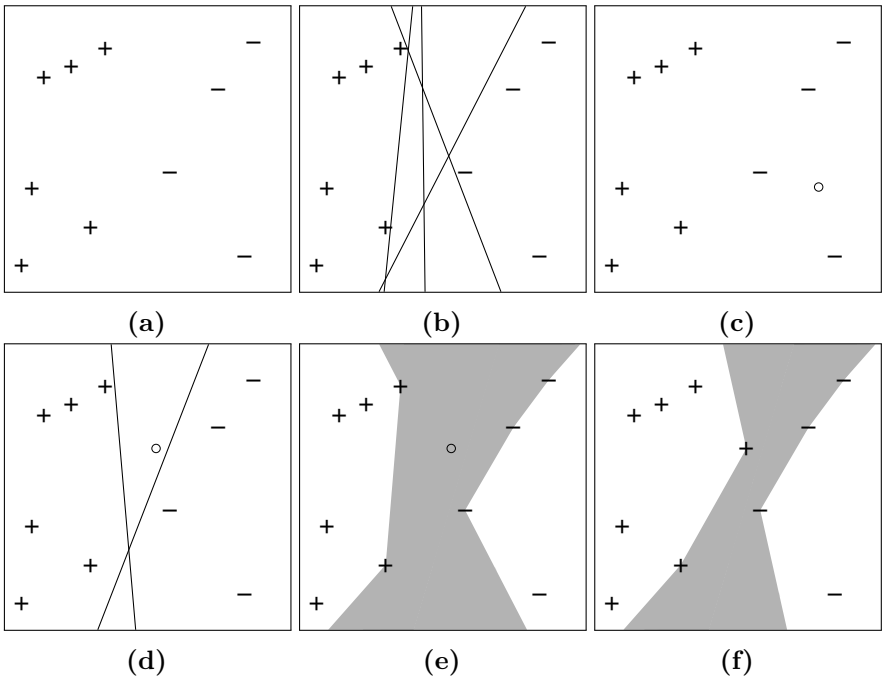


Figure 1.1: An illustration of the concepts involved in disagreement-based active learning, in the context of learning a linear separator in 2 dimensions.

In this particular case, note that *all* of the lines separating the observed "+" points from the observed "−" points have this new point on the "−" side of the line. Since we know the target separator is among these lines, we can conclude that the correct label of this new point is "−". The important detail here is that *we did not need to observe the correct label in order to deduce its value.*

On the other hand, what if instead we are given the unlabeled point depicted in Figure 1.1d? In this case, there is some line that correctly separates the other points while including this new point on the "−" side, and there is another line that correctly separates the other points while including this new point on the "+" side. So we are unable to deduce the correct label of this point based only on the information already available. The disagreement-based active learning strategy is characterized by the fact that it will *request* the value of

the label (from the expert/oracle) whenever (and only whenever) this is the case. Indeed, for this data set, the disagreement-based strategy would make a label request when presented with any unlabeled point in the shaded region in Figure 1.1e: namely, the set of points such that there is some disagreement among the separators consistent with the observed labels. This set is referred to as the *region of disagreement* (or region of uncertainty).

Since the disagreement-based active learning strategy requests the label of a sample only if it is in the region of disagreement, the analysis of the label complexity of this strategy hinges on understanding the probability a new sample will be inside the region of disagreement. In particular, we will be interested in how this probability behaves as a function of the number of observed data points. The good news is that often (though not always) this probability decreases as the data set grows. For instance, suppose, in response to our request, we are told that the label of the new point in Figure 1.1d is "+". If we then add this point to the data set, the *new* region of disagreement becomes the shaded region in Figure 1.1f, which is a significant reduction compared to the region in Figure 1.1e (e.g., under a uniform probability measure within the figure). In the next chapter, we will introduce a quantity called the *disagreement coefficient*, which helps us to characterize the *rate of decrease* of the probability of getting a point in the region of disagreement.

One of the most remarkable facts about this idea is that it is fully *general*, in the sense that the exact same principle can be used in combination with *any* type of classifier. For instance, consider instead the problem of learning an *axis-aligned rectangle* in the 2-dimensional plane: that is, the label of each point is "+" if the point is contained inside an (unknown) rectangle $[a_1, b_1] \times [a_2, b_2]$ in the plane, and is "−" if the point is outside this rectangle. Suppose we have obtained a data set as depicted in Figure 1.2a. A few of the rectangles consistent with these labels are depicted in Figure 1.2b (again, there are in fact an infinite number of consistent rectangles). The region of disagreement is then depicted as the shaded region in Figure 1.2c. Thus, if we are given a new sample outside this shaded region, we can deduce its la-

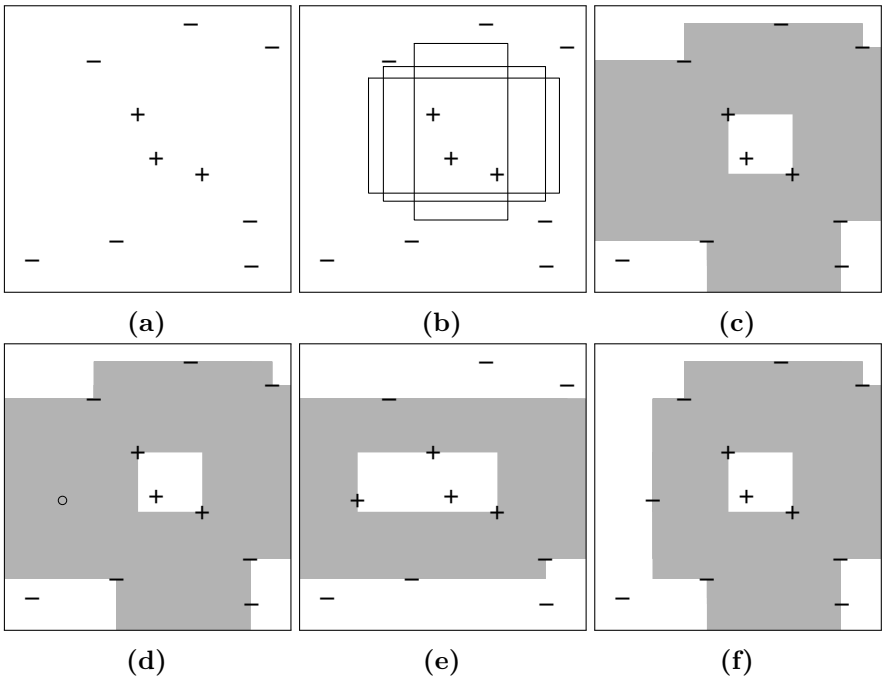
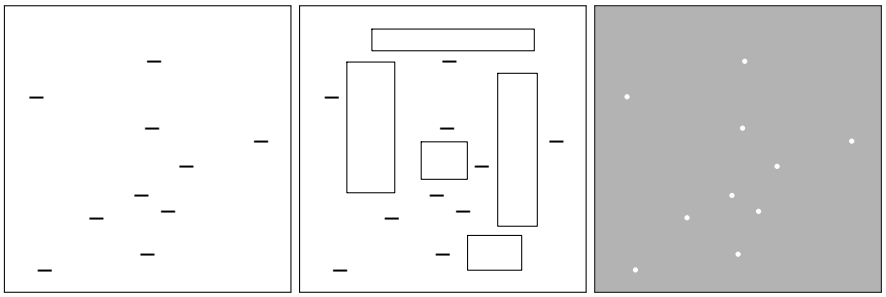


Figure 1.2: The same core idea of disagreement-based active learning can be applied with any type of classifier. Here we illustrate these concepts in the context of learning an axis-aligned rectangle in 2 dimensions.

bel without requesting its value; in the interior unshaded region, the deduced label would be "+", while in the exterior unshaded region, the deduced label would be "−". Again, the disagreement-based active learning strategy would request the label of a new point if and only if it is inside the shaded region. As before, given the requested label of a point in the shaded region, adding this labeled point to the data set would cause a reduction in the region of disagreement. For instance, for the new point marked "∘" in Figure 1.2d, if we are told the correct label is "+", upon adding this point to the data set, the new region of disagreement would be the shaded region depicted in Figure 1.2e; on the other hand, if we are told the correct label is "−", the new region of disagreement would be the shaded region depicted in Figure 1.2f.

In both of the scenarios described above, requesting the labels of



(a) (b) (c)

Figure 1.3: In the context of learning an axis-aligned rectangle, if all of the observed labels are "−", *every* point not in the data set is contained in the region of disagreement.

points in the region of disagreement resulted in a significant decrease in the region of disagreement. These would be considered *favorable* scenarios for disagreement-based active learning. However, we are not always so fortunate. For instance, consider again the scenario where a point is labeled "+" iff it is contained inside an unknown rectangle $[a_1, b_1] \times [a_2, b_2]$ in the plane, but this time suppose the data set observed so far is as depicted in Figure 1.3a. Note that all of the points in this data set are labeled "−". In this case, *every* rectangle that does not contain any of these data points would be consistent with their labels; a few such rectangles are depicted in Figure 1.3b. It should be clear that this is a very different kind of scenario from the prevous two. In particular, for every point (x_1, x_2) in the plane that is not among the few observed samples, the rectangle $[x_1, x_1] \times [x_2, x_2]$ containing *only this point* is consistent with all of the observed labels. Since this is true of *every* point not among the observed samples, the region of disagreement is the *entire space*, minus the few points in the data set; this is represented by the shaded region in Figure 1.3c. Thus, if we are given a new point that is not equal to one we have already observed the label of, the disagreement-based strategy will request its label. If, in response, we are told that the label is "−", then the region of disagreement is reduced by *only this single point*. In particular, if the probability distribution is non-atomic, then no matter how many

samples labeled "−" we observe, the probability in the region of disagreement will always equal 1, and therefore *does not decrease*. Thus, if the unknown target rectangle has *zero* probability inside, then this situation will continue indefinitely (with probability 1), requesting every label and never reducing the probability in the region of disagreement.

The distinction raised by contrasting these two kinds of scenarios is fundamental to the active learning problem. In the chapters below, we will be highly interested in discussions of general conditions that distinguish between problems where the probability in the region of disagreement decreases (and approaches zero) and those where it does not. In the former case, we will be further interested in understanding the rates of decrease. With this understanding in hand, we are then able to describe the label complexities achieved by certain disagreement-based active learning algorithms *abstractly*. Various specific scenarios, such as those described above, can then be studied straightforwardly as special cases of the general analysis.

References

K. S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields*, 75:379–423, 1987.

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Machine Learning*, 30:31–56, 1998.

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

M.-F. Balcan and S. Hanneke. Robust interactive learning. In *Proceedings of the 25$^{\text{th}}$ Conference on Learning Theory*, 2012.

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the 20th Conference on Learning Theory*, 2007.

M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *Proceedings of the* 21st *Conference on Learning Theory*, 2008.

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

M.-F. Balcan, S. Hanneke, and J. Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2–3):111–139, 2010.

J. L. Balcázar, J. Castro, and D. Guijarro. A new abstract combinatorial dimension for exact learning via queries. *Journal of Computer and System Sciences*, 64(1):2–21, 2002.

P. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems 23*, 2010.

A. Beygelzimer, D. Hsu, N. Karampatziakis, J. Langford, and T. Zhang. Efficient active learning. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

V. I. Bogachev. *Gaussian Measures*. American Mathematical Society, Mathematical Surveys and Monographs, Book 62, 1998.

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

R.M. Castro and R.D. Nowak. Upper and lower error bounds for active learning. In *The 44th Annual Allerton Conference on Communication, Control and Computing*, 2006.

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

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334, 1965.

S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*, 2005.

S. Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 412: 1767–1781, 2011.

S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the 18th Conference on Learning Theory*, 2005.

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*, 2007.

S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, 10:281–299, 2009.

O. Dekel, C. Gentile, and K. Sridharan. Robust selective sampling from single and multiple teachers. In *Proceedings of the* 23$^{\text{rd}}$ *Conference on Learning Theory*, 2010.

O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, To Appear, 2012.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, Inc., 1996.

R. M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 15(4):1306–1326, 1987.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

B. B. Eisenberg. *On the Sample Complexity of PAC-Learning using Random and Chosen Examples*. PhD thesis, Massachusetts Institute of Technology, 1992.

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

R. El-Yaniv and Y. Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13:255–279, 2012.

V. Feldman, P. Gopalan, S. Khot, and A.K. Ponnuswami. On agnostic learning of parities, monomials and halfspaces. *SIAM Journal on Computing*, 39(2): 606–645, 2009.

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

E. Friedman. Active learning for smooth problems. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50:20–31, 1995.

V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

S. Hanneke. Teaching dimension and the complexity of active learning. In *Proceedings of the 20^{th} Conference on Learning Theory*, 2007a.

S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24^{th} International Conference on Machine Learning*, 2007b.

S. Hanneke. Adaptive rates of convergence in active learning. In *Proceedings of the* 22^{nd} *Conference on Learning Theory*, 2009a.

S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009b.

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

S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13(5):1469–1587, 2012.

S. Hanneke. Theory of Active Learning, 2014. URL `http://www.stat.cmu.edu/~shanneke`.

S. Hanneke and L. Yang. Negative results for active learning with convex losses. In *Proceedings of the* 13^{th} *International Conference on Artificial Intelligence and Statistics*, 2010.

S. Hanneke and L. Yang. Surrogate losses in passive and active learning. *arXiv:1207.3772*, 2012.

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

D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory A*, 69:217–232, 1995.

D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115:248–292, 1994.

T. Hegedüs. Generalized teaching dimension and the query complexity of learning. In *Proceedings of the 8th Conference on Computational Learning Theory*, 1995.

D. Helmbold, R. Sloan, and M. Warmuth. Learning nested differences of intersection-closed concept classes. *Machine Learning*, 5:165–196, 1990.

D. Hsu. *Algorithms for Active Learning*. PhD thesis, Department of Computer Science and Engineering, School of Engineering, University of California, San Diego, 2010.

M. Kääriäinen. Active learning in the non-realizable case. In *Proceedings of the 17th International Conference on Algorithmic Learning Theory*, 2006.

A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46^{th} Annual IEEE Symposium on Foundations of Computer Science*, 2005.

M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

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

S. R. Kulkarni. On metric entropy, Vapnik-Chervonenkis dimension, and learnability for a class of distributions. Technical Report CICS-P-160, Center for Intelligent Control Systems, 1989.

S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11:23–35, 1993.

S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

P. M. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6): 1556–1559, 1995.

S. Mahalanabis. A note on active learning for smooth problems. arXiv*:1103.3095*, 2011.

S. Mahalanabis. *Subset and Sample Selection for Graphical Models: Gaussian Processes, Ising Models and Gaussian Mixture Models*. PhD thesis, Department of Computer Science, Edmund A. Hajim School of Engineering & Applied Sciences, University of Rochester, Rochester, New York, 2012.

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

P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48:1977–1991, 2002.

S. Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13(1):67–90, 2012.

J. R. Munkres. *Topology*. Prentice Hall, 2^{nd} edition, 2000.

R. D. Nowak. Generalized binary search. In *Proceedings of the 46*$^{\text{th}}$ *Allerton Conference on Communication, Control, and Computing*, 2008.

R. D. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12), 2011.

D. Pollard. *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2, Institute of Mathematical Statistics and American Statistical Association, 1990.

M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems 24*, 2011.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

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

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

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

A. van der Vaart and J. A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

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

H. S. Witsenhausen. A counterexample in stochastic optimum control. *SIAM Journal of Control*, 6(1):131–147, 1968.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004.