Mathematical Theories of Interaction with Oracles

Liu Yang

2013

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Avrim Blum
Manuel Blum
Jaime Carbonell
Sanjoy Dasgupta
Yishay Mansour
Joel Spencer

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

This thesis is dedicated to all Mathematicians.

Acknowledgments

I would like to thank my advisor Avrim Blum for so many stimulating discussions (research problems and other fun math problems), for the inspiration I experienced during our discussions, for his amazingly accurate-with-high-probability sense of the directions that are worth trying, and for the many valuable bits of feedback and advice he has provided me. I also thank my other advisor Jaime Carbonell for always being supportive and encouraging me to push on with one problem after another. I am grateful to Manuel Blum for so many ingenious discussions all through these years when I am at CMU, which have broadened my mind, and given me a great taste of research problems and a faith in the ability of Mathematics to uncover interesting and mysterious truths, such as the nature of consciousness. I appreciate the exhilarating experience of working with Yishay Mansour on an algorithmic economics problem; through these interactions, I have learned many insights about axiomatic approaches to algorithmic economics.

One of my great experiences has been interacting with many wonderful mathematicians. I thank Ryan O'Donnell for input on my research on learning DNF, and insights on the analysis of boolean functions. I appreciate discussions with Steven Rudich on interactive proof systems, and for his counseling on Fourier techniques; he has also helped sharpen my skills of giving good talks and lectures. I thank Venkatesan Guruswami for discussions on information theory and coding theory related to my work in Bayesian active learning; I also highly enjoyed his complexity theory class. I want to thank Tuomas Sandholm for sharing his knowledge of Bayesian auction design. I thank Anupam Gupta for discussions on approximation algorithms. I would also like to thank all the other faculty that I've interacted with in my time at CMU. Thanks especially to my co-author Silvio Micali for extending my philosophical and implementational insights on auction design. I thank Shafi Goldwasser for encouragement on my work in property testing and computational learning theory. I thank Leslie Valiant for input on my project on learning DNF with representation-specific queries.

There are also several mathematicians who, though our interactions have been only brief, have made a lasting impact on my mathematical perspective. I am grateful for the wonderful and stimulating discussion I had with Alan Frieze on combinatorics. I appreciate the one sentence of advice from John Nash when I happened to be at Princeton for a summer workshop. I am grateful to Scott Aaronson and Avi Wigderson for a few email conversations on interactive proof systems with restricted provers, which is a project I am actively pursuing. I also thank all the theorists I met in conferences, and the many friends and peers that made my time as a graduate student quite enjoyable, including Eric Blais and Paul Valiant. Finally, I want to cite Fan Chung Graham's advice for grad students "Coauthorship is a closer relationship than friendship." Yes, indeed, the co-authorship with all my collaborators is to be cherished year after year.

Contents

Chapter 1

Summary

The key insight underlying this thesis is that the right kind of interaction is the key to making the intractable tractable. This work specifically investigates this insight in the context of learning theory. While much of the learning theory literature has traditionally focused on protocols that are either non-interactive or involving unrealistically strong forms of interaction, there have recently been several exciting advances in the design and analysis of methods for realistic interactive learning protocols.

Perhaps one of the most interesting of these is *active learning*. In active learning, a learning algorithm is given access to a large pool of unlabeled examples, and is allowed to sequentially request their labels so as to learn how to accurately predict the labels of new examples. This thesis contains a number of interesting advances in our understanding of the capabilities of active learning methods. Specifically, I summarize the main contributions below.

1.1 Bayesian Active Learning

While most of the recent advances in our understanding of active learning have focused on the traditional PAC model (or noisy variants thereof), similar advnaces specific to the Bayesian learning setting have largely been lacking. Specifically, suppose that in addition to the data itself, the

learner additionally has access to a *prior* distribution for the target function, and we are interested in achieving a guarantee of low expected error rate, where the expectation is over both the draw of the data *and* the draw of the target concept from the given prior. This setting has been studied in depth for the passive learning protocol, but aside from the well-known work on the query-by-committee algorithm, little was known about this setting for the active learning protocol. This lack of knowledge is particularly troubling in light of the fact that most of the active learning methods used in practice have Bayesian interpretations, selecting their label requests based on Bayesian notions such as label entropy, expected error reduction, or reduction in the total probability mass of the version space.

1.1.1 Arbitrary Binary-Valued Queries

In this thesis, we present work that makes progress in understanding the Bayesian active learning setting. To begin, we study the most basic question: how many queries are necessary if we are able to ask *arbitrary* binary-valued queries. While label requests are only a special type of binary-valued query, a general lower bound for arbitrary binary-valued queries will also hold for label request queries, and thus provides a lower bound on the intrinsic query complexity of the learning problem. Not surprisingly, we find that the number of binary-valued queries necessary for learning is characterized by a kind of entropy quantity: namely, the entropy of the Voronoi partition induced by a maximal ϵ-packing.

1.1.2 Self-Verifying Active Learning

Our next contribution is a study of a special type of active learning, characterized by the stopping-criterion used in the learning algorithm. Specifically, consider a protocol in which the input to the active learning algorithm is the desired error rate guarantee ϵ, and the algorithm then makes a number of queries and then halts. For the algorithm to be considered "correct", it must have the guarantee that the expected error rate of the classifier it produces after halting is at most

the value of ϵ provided as input. We refer to this family of algorithms as *self-verifying*. The label complexity of learning in this protocol is generally higher than in some other protocols (e.g., budget-based), since the algorithm must not only *find* a classifier with good error rate, but must also somehow be *self-aware* of the fact that it has found such a good classifier. Indeed, it is known that prior-independent self-verifying algorithms may often have label complexities no better than that of passive learning, which is $\Theta(1/\epsilon)$ for VC classes. However, we prove that in Bayesian active learning, for any VC class and prior, there is a prior-dependent method that always achieves an expected label complexity that is $o(1/\epsilon)$. Thus, this represents a concrete result on the advantages of having access to the target's prior distribution.

1.2   Active Testing

One of the major challenges facing active learning is that of model selection. Specifically, given a number of hypothesis classes, how does one decide which one to use? In passive learning, the solution is simple: try them all, and then pick from among the resulting hypotheses using cross-validation. But such solutions are not available to active learning, since the methods tailored to each hypothesis class will generally make very different label requests, so that the label complexity of producing a hypothesis from all of the classes is close to the sum of their individual label complexities.

Thus, to avoid this problem, there is a need for procedures that quickly dermine whether the target concept is within (or approximated by) a given concept class, by asking a much smaller number of label requests than required for *learning* with that class: that is, for *testing* methods that operate in the active learning protocol, which we therefore refer to as *active testing*. This way, we can simply go through each class and test whether the target is in the class or not, and only run the full learning method on some simplest class that passes the test. The questions then become how many fewer queries are required for testing compared to learning, as this quantifies the savings from using this approach. Following the traditional literature on property testing,

the primary focus of such an analysis is on the dependence of the query complexity on the VC dimension of the hypothesis class being tested. Since learning typically required a number of queries linear in the VC dimension, a sublinear dependence is considered an improvement, while a query complexity independent of the VC dimension is considered superb.

There is much existing literature on property testing. However, the standard model of property testing makes use of *membership queries*, which are effectively label requests for feature vectors of our own construction, rather than feature vectors from a given polynomial-sized sample of unlabeled examples from the data distribution. Such methods are unrealistic for our model selection purposes, since it is well-known in the machine learning community that the feature vectors constructed by membership queries are often unintelligible by the human experts charged with labeling the examples. However, the results from this literature on membership queries do provide us a useful additional reference point, since we are certain that the query complexity of active testing is no smaller than that of testing with membership queries, and no larger than that of testing from random labeled examples (passive testing).

In our work on active testing, we study a number of interesting concept classes, and find that in some cases the query complexity is nearly the same as that of testing with membership queries, while other times it is closer to that of passive testing. However, in most (though not all) cases, we do find that the query complexity of active testing is significantly smaller than that of active *learning*, so that this approach to model selection can indeed be quite effective at reducing the total query complexity.

1.3   Theory of Transfer Learning

Given the positive results mentioned above on the advantages of active learning with access to the target's prior distribution, the next natural quesiton is, "How does one gain access to the target's prior distribution?" Traditionally, there have been a variety of answers to this question given by the Bayesian Statistics community, ranging from subjective beliefs, to computationally-

motivated assumptions, to estimation. Perhaps one of the most appealing, from a practical perspective, it the Empirical Bayes perspective, which says that we gain access to an approximation of the prior based on analysis of past experience. In the learning context, this idea of gaining insights for a new learning problem, based on experience with past learning problems, goes by the name *Transfer Learning*. The specific model of transfer learning relevant to this Empirical Bayes setting is the following. We suppose that we are tasked with a sequence of T learning problems, or *tasks*. For each task, the unlabeled data are sampled i.i.d. according to some distribution $\mathcal{D}$, independently across the tasks. Furthermore, for each task the target function is sampled according to some prior distribution π, again independently across tasks. We then approach each task as usual, making a number of label requests and then halting with guaranteed expected error rate at most ϵ. The hope is that, after solving a number of learning problems $t < T$, the label complexity of solving task $t + 1$ should be smaller than that of solving the first task, due to gaining some information about the distribution π.

The challenge in this problem is that we do not get direct observations of the target functions from each task. Rather, we may only observe a small number of labeled examples. So the question is how to extract useful information about π from these limited observations. This situation is further complicated by the fact that we are interested in minimizing the number of samples per-task, and that the active learning method's queries might be highly task-specific. Indeed, in many transfer learning settings, each task is approached by a different agent, who may be non-altruistic with respect to the other agents; thus, she may be unwilling to make very many additional label requests merely to aid the learners that will solve future tasks.

In our work, we show that it is possible to gain benefits from transfer learning, while limiting the number of additional queries (other than those used directly for learning) required from each task. Specifically, we use a number of extra queries per task equal the VC dimension of the concept class. Using these queries, we are able to consistently estimate π, assuming only that it resides in a known totally bounded class of distributions. We are then able to use this esti-

mate in the context of a prior-dependent learning method to asymptotically achieve an average label complexity equal to that of learning with *direct* knowledge of π. Thus, we have realized the aforementioned benefits of having knowledge of the target's prior, including the guaranteed $o(1/\epsilon)$ expected label complexity for self-verifying active learning. We further show that no method taking fewer than VC dimension number of samples per task can match this guarantee at this level of generality.

Interestingly, under smoothness conditions on π, we also provide explicit bounds on the *rate* of convergence of our estimator to π, and we additionally derive lower bounds on the minimax rate of convergence. This has implications for non-asymptotic guarantees on the benefits of transfer learning.

We also extend these results to real-valued functions, where the VC dimension is replaced by the pseudo-dimension of the function class. In addition to transfer learning, we also find that this technique for estimating a prior distribution over real-valued functions has applications to the preference elicitation problem in a certain type of combinatorial auction.

1.4   Active Learning with Drifting Distributions and Targets

In addition to the work on Bayesian active learning, I have additionally studied the setting of active learning without access to a prior. Work in this area is presently more mature, so that there are known methods that are robust to noise, and have well-understood label complexities. However, all of the previous theoretical work on active learning supposed the data were sampled i.i.d. from some fixed (though unknown) distribution. But many realistic applications of active learning involve distributions that change over time, so that we require some understanding of how active learning methods behave under drifting distributions.

In my work on this topic, I study a model of distribution drift in which the conditional distribution of label given features remains fixed (i.e., no target drift), while the marginal distribution over the feature vectors can change arbitrarily within a given totally bounded family of distribu-

tions from one observation to the next. I then analyze a stream-based active learning setting, in which the learner is at each time required to make a prediction for the label of a new example, and then decide whether to request the label or not. We are then interested in the expected number of mistakes and number of label requests, as a function of how many data points have been observed.

Interestingly, I find that even with such drifting distributions, it is still possible to guarantee a number of mistakes on par with fully-supervised learning, while only requesting a sublinear number of labels, as long as the disagreement coefficient is sublinear in the reciprocal of its argument under all distributions in the given family. I prove this, both under the realizable case, and under Tsybakov noise conditions. I further provide a more detailed analysis of the frequency of label requests and mistakes, as a function of the Tsybakov noise parameters, the supremum of the disagreement coefficient over the given family of distributions, and the covering numbers of the family of distributions. To complement this, I also provide lower bounds on the number of label requests required of any active learning method whose number of mistakes is on par with the optimal performance of fully-supervised learning.

We have also studied the related problem of active learning with a drifting target concept, in which the target function itself changes over time. In this setting, the distribution over unlabeled samples remains fixed, while the function providing labels changes over time at a specified rate. We then express bounds on the expected number of mistakes and queries, as a function of this rate of change and the number of samples.

In any learning context, the problem of efficient learning in the presence of noise is a constant challenge. Toward addressing this challenge, we have proposed an active learning algorithm that makes use of a convex surrogate loss function, in place of the 0-1 loss, while still providing guarantees on the obtained error rate (under the 0-1 loss) and number of queries made in the active learning context, under the assumption that the surrogate loss is classification-calibrated, and the minimizer of the surrogate loss resides in the function class used by the algorithm.

1.5 Efficiently Learning DNF with Representation-Specific Queries

In addition to the basic active learning protocol, based on label requests, we have also studied an interesting new type of learning protocol, in which the algorithm is allowed queries regarding specific aspects of the *representation* of the target function. This setting is motivated by applications in which there are essentially sub-labels for the examples, which may be difficult for an expert to explicitly produce, but for which they can easily recognize commonality. For instance, in fraud detection, we may be able to ask an expert whether two given examples of fraudulent transactions are representative of the same *type* of fraud.

To study this idea in formality, we specifically look at the classic problem of efficiently learning a DNF formula. Certain variants of this problem are known to be NP-Hard if we are permitted only labeled data (e.g., proper learning), and there are no known efficient methods for the general problem of learning DNF, even with membership queries. In fact, under the uniform distribution, there are no such general results known even for the problem of learning monotone DNF from labeled data alone. Thus, there is a real need for new ideas to approach the problem of learning DNF if the class of DNF functions is to be used for practical applications.

In our work, we suppose access to a polynomial-sized sample of labeled examples, and for any pair of positive examples from that sample, we allow queries of the type, "Do these two examples satisfy a term in common in the target DNF?" It turns out that the problem of learning arbitrary DNF under arbitrary distributions is no easier with this type of query than with labeled examples alone. However, using queries of this type, we are able to efficiently learn several interesting sub-families of DNF, including solving some problems known to be NP-Hard from labeled data alone (properly learning 2-term DNF). Additionally, under the uniform distribution, we find many more interesting families of DNF that are efficient learnable with queries of this type, including the well-studied family of $O(\log(n))$-juntas, and any DNF for which each variable appears in at most $O(\log(n))$ terms.

We further study several generalizations of this type of query. In particular, if we allow the

algorithm to ask "How many terms do these two examples satisfy in common in the target DNF?" then we can significantly broaden the collection of subfamilies of DNF that are efficiently learnable. In particular, $O(\log(n))$-juntas become efficiently learnable under arbitrary distributions, as does the family of DNF with $O(\log(n))$ terms.

With a further strengthening to allow the query to involve an arbitrary number of examples, rather than just two, we find we can efficiently (properly) learn an arbitrary DNF under an arbitrary distribution. This is also the case if we restrict to just two examples in the query, but we allow the algorithm to construct the feature vectors for those two examples, rather than selecting them from a polynomial-sized sample.

Overall, we feel this is an important topic, in that it makes real progress on the practically-important problem of efficiently learning DNF, which has otherwise been essentially stagnant for a number of years.

1.6   Online Allocation with Economies of Scale

In addition to all of the above work on computational learning theory, this dissertation also includes work on allocations problems in which the cost of allocating each additional copy of a good is decreasing in the number of copies already allocated. This model captures the natural economies of scale that arise in many real-world contexts. In this context, we derive methods capable of allocating goods to a set of customers in a unit-demand setting, while achieving near-optimal cost guarantees. We study this problem both in an offline setting, in which all of the customer valuation functions are known in advance, and also in a type of online setting, in which the customers arrive one-at-a-time, so that we do not know in advance what their valuation functions will be. In the online variant of the problem, working under the assumption that the valuation functions are i.i.d. samples, we make use of generalization guarantees from statistical learning theory, in combination to the algorithmic solutions to the offline problem, to obtain the approximation guarantees.

Chapter 2

Active Testing

Abstract

[1] One of the motivations for property testing of boolean functions is the idea that testing can serve as a preprocessing step before learning. However, in most machine learning applications, the ability to query functions at arbitrary points in the input space is considered highly unrealistic. Instead, the dominant query paradigm in applied machine learning, called *active learning*, is one where the algorithm may ask for examples to be labeled, but *only from among those that exist in nature*. That is, the algorithm may make a polynomial number of draws from the underlying distribution D and then query for labels, but only of points in its sample. In this work, we bring this well-studied model in learning to the domain of *testing*, calling it *active testing*.

We show that for a number of important properties, testing can still yield substantial benefits in this setting. This includes testing unions of intervals, testing linear separators, and testing various assumptions used in semi-supervised learning. For example, we show that testing unions of d intervals can be done with $O(1)$ label requests in our setting, whereas it is known to require $\Omega(\sqrt{d})$ labeled examples for passive testing (where the algorithm must pay for labels on *every* example drawn from D) and $\Omega(d)$ for learning. In fact, our results for testing unions of intervals

---

[1]Joint work with Maria-Florina Balcan, Eric Blais, and Avrim Blum.

also yield improvements on prior work in both the membership query model (where any point in the domain can be queried) and the passive testing model [Kearns and Ron, 2000] as well. In the case of testing linear separators in R^n, we show that both active and passive testing can be done with $O(\sqrt{n})$ queries, substantially less than the $\Omega(n)$ needed for learning and also yielding a new upper bound for the passive testing model. We also show a general combination result that any disjoint union of testable properties remains testable in the active testing model, a feature that does not hold for passive testing.

In addition to these specific results, we also develop a general notion of the *testing dimension* of a given property with respect to a given distribution. We show this dimension characterizes (up to constant factors) the intrinsic number of label requests needed to test that property; we do this for both the active and passive testing models. We then use this dimension to prove a number of lower bounds. For instance, interestingly, one case where we show active testing does *not* help is for dictator functions, where we give $\Omega(\log n)$ lower bounds that match the upper bounds for learning this class.

Our results show that testing can be a powerful tool in realistic models for learning, and further that active testing exhibits an interesting and rich structure. Our work in addition develops new characterizations of common function classes that may be of independent interest.

2.1 Introduction

One of the motivations for property testing of boolean functions is the idea that testing can serve as a preprocessing step before learning – to determine whether learning with a given hypothesis class is worthwhile [Goldreich, Goldwasser, and Ron, 1998]. Indeed, query-efficient testers have been designed for many common hypothesis classes in machine learning such as linear threshold functions [Matulef, O'Donnell, Rubinfeld, and Servedio, 2009], unions of intervals [Kearns and Ron, 2000], juntas [Blais, 2009, Fischer, Kindler, Ron, Safra, and Samorodnitsky, 2004], DNFs [Diakonikolas, Lee, Matulef, Onak, Rubinfeld, Servedio, and Wan, 2007], and decision

trees [Diakonikolas, Lee, Matulef, Onak, Rubinfeld, Servedio, and Wan, 2007]. (See Ron's survey [Ron, 2008] for much more on the connection between learning and property testing.)

Most property testing algorithms, however, rely on the ability to query functions on arbitrary points – an assumption that is unrealistic in most machine learning applications. For example, in classifying documents by topic, while selecting an existing document on the web and asking a user "is this about sports or business?" may make perfect sense, taking an existing sports document (represented in R^n as a vector of word-counts), corrupting a random fraction of the entries, and asking "is this still about sports?" does not. Early experiments yielded similar failures for membership-query learning algorithms in vision applications when asking human users about corrupted images [Baum and Lang, 1993]. As a result, the dominant query paradigm in machine learning has instead been the model of *active learning* where the algorithm may query for labels of examples of its choosing, but *only among those that exist in nature* [Balcan, Beygelzimer, and Langford, 2006, Balcan, Broder, and Zhang, 2007a, Balcan, Hanneke, and Wortman, 2008, Beygelzimer, Dasgupta, and Langford, 2009, Castro and Nowak, 2007, Cohn, Atlas, and Ladner, 1994a, Dasgupta, 2005, Dasgupta, Hsu, and Monteleoni, 2007b, Hanneke, 2007a, Seung, Opper, and Sompolinsky, 1992, Tong and Koller., 2001].

In this work, we bring this well-studied model in learning to the domain of *testing*. In particular, we assume that as in active learning, our algorithm can make a polynomial number of draws of *unlabeled examples* from the underlying distribution D (these unlabeled examples are viewed as cheap), and then can make a small number of label queries but *only* over the unlabeled examples drawn (these label queries are viewed as expensive). The question we ask is whether testing in this setting is sufficient to still yield significant benefit in terms of label requests over the number of labeled examples needed for learning.

What we show is that for a number of interesting properties relevant to learning, this capability indeed allows for a substantial reduction in the number of labels required. This includes testing unions of intervals, testing linear separators, and testing various assumptions about the

separation of data used in semi-supervised learning. For example, we show that testing unions of d intervals can be done with $O(1)$ label requests in our setting, whereas it is known to require $\Omega(\sqrt{d})$ labeled examples for passive testing (where the algorithm must pay for labels on *every* example drawn from D) and $\Omega(d)$ for learning. In the case of testing linear separators in R^n, we show that both active and passive testing can be done with $O(\sqrt{n})$ queries, substantially less than the $\Omega(n)$ needed for learning and also yielding a new upper bound for the passive testing model as well. These results use a generalization of Arcones Theorem on the concentration of U-statistics. For the case of unions of intervals, our results even improve on prior work in the membership query and passive models of testing [Kearns and Ron, 2000], and are based on a characterization of this class in terms of noise sensitivity that may be of independent interest. We also show that any disjoint union of testable properties remains testable in the active testing model, allowing one to build testable properties out of simpler components; this is a feature that does not hold for passive testing.

In addition to the above results, we also develop a general notion of the *testing dimension* of a given property with respect to a given distribution. We show this dimension characterizes (up to constant factors) the intrinsic number of label requests needed to test that property; we do this for both passive and active testing models. We then make use of this notion of dimension to prove a number of lower bounds. For instance, one interesting case where we show active testing does *not* help is for dictator functions, a classic property where membership queries can allow testing with $O(1)$ label requests, but where we show active testing requires $\Omega(\log n)$ labels, matching the bounds for learning.

Our results show that a number of important properties for learning can be tested with a small number of label requests in a realistic model, and furthermore that active testing exhibits an interesting and rich structure. We further point out that unlike the case of passive learning, there are no known strong Structural Risk Minimization bounds for active learning, which makes the use of testing in this setting even more compelling.[2] Our techniques are quite different from

[2]In passive learning, if one has a collection of algorithms or hypothesis classes to try, there is little advantage

13

those used in the active learning literature.

2.1.1 The Active Property Testing Model

Before discussing our results in more detail, let us first introduce the model of active testing. A *property* $\mathcal{P}$ of boolean functions is simply a subset of all boolean functions. We will also refer to properties as *classes* of functions. The *distance* of a function f to the property $\mathcal{P}$ over a distribution D on the domain of the function is $dist_D(f, \mathcal{P}) := \min_{g \in \mathcal{P}} \Pr_{x \sim D}[f(x) \neq g(x)]$. A *tester* for $\mathcal{P}$ is a randomized algorithm that must distinguish (with high probability) between functions in $\mathcal{P}$ and functions that are far from $\mathcal{P}$. In the standard property testing model introduced by Rubinfeld and Sudan [Rubinfeld and Sudan, 1996], a tester is allowed to query the value of the function on any input in order to make this decision. We consider instead a model in which we add restrictions to the possible queries:

Definition 2.1 (Property tester)**.** *An s-sample, q-query ϵ-tester for $\mathcal{P}$ over the distribution D is a randomized algorithm A that draws s samples from D, sequentially queries for the value of f on q of those samples, and then*

1. *Accepts w.p. at least $\frac{2}{3}$ when $f \in \mathcal{P}$, and*

2. *Rejects w.p. at least $\frac{2}{3}$ when $dist_D(f, \mathcal{P}) \geq \epsilon$.*

We will use the terms "label request" and "query" interchangeably. Definition 2.1 coincides with the standard definition of property testing when the number of samples is unlimited and the distribution's support covers the entire domain. In the other extreme case where we fix $q = s$, our definition then corresponds to the *passive testing* model, where the inputs queried by the tester are sampled from the distribution. Finally, by setting s to be polynomial in some appropriate measure of the input domain, we obtain the *active testing* model that is the focus of this paper:

asymptotically to being told which of these is best in advance, since one can simply apply all of them and use an appropriate union bound. In contrast, this is much less clear for active learning algorithms that each might ask for labels on different examples.

14

Definition 2.2 (Active tester). *A randomized algorithm is a q-query active ϵ-tester for $\mathcal{P} \subseteq \{0,1\}^n \to \{0,1\}$ over D if it is a $\mathrm{poly}(n)$-sample, q-query ϵ-tester for $\mathcal{P}$ over D.*

Remark 2.1. *We emphasize that the name* active tester *is chosen to reflect the connection with active learning. It is* not *meant to imply that this model of testing is somehow "more active" than the standard property testing model.*

In some cases, the domain of our functions is not $\{0,1\}^n$. In those cases, we require s to be polynomial in some other appropriate measure of complexity that we specify explicitly.

Note that in Definition 2.1, since we do not have direct membership query access (at arbitrary points), our tester must accept w.p. at least $\frac{2}{3}$ when f is such that $dist_D(f, \mathcal{P}) = 0$, even if f does not satisfy $\mathcal{P}$ over the entire input space. This, in fact, is one crucial difference between our model and the *distribution-free* testing model introduced by Halevy and Kushilevitz [Halevy and Kushilevitz, 2007] and further studied in [Dolev and Ron, 2010, Glasner and Servedio, 2009, Halevy and Kushilevitz, 2004, 2005]. In the distribution-free model, the tester can sample inputs from some unknown distribution and can query the target function on *any* input of its choosing. It must then distinguish between the case where $f \in \mathcal{P}$ from the case where f is far from the property over the distribution. Most testers in this model strongly rely on the ability to query any input[3] and, therefore, these algorithms are not valid active testers.

In fact, the case of dictator functions, functions $f : \{0,1\}^n \to \{0,1\}$ such that $f(x) = x_i$ for some $i \in [n]$, helps to illustrate the distinction between active testing and the standard (membership query) testing model. The dictatorship property is testable with $O(1)$ membership queries [Bellare, Goldreich, and Sudan, 1998, Parnas, Ron, and Samorodnitsky, 2003]. In contrast, with active testing, the query complexity is the same as needed for learning:

Theorem 2.3. *Active testing of dictatorships under the uniform distribution requires $\Omega(\log n)$ queries. This holds even for distinguishing dictators from random functions.*

---

[3]Indeed, Halevy and Kushilevitz's original motivation for introducing the model was to better model PAC learning in the *membership query* model [Halevy and Kushilevitz, 2007].

This result, which we prove in Section 2.5.1 as an application of the active testing dimension defined in Section 2.5, points out that the constraints imposed by active testing present real challenges. Nonetheless, we show that for a number of interesting properties we can indeed perform active testing with substantially fewer queries than needed for learning or passive testing. In some cases, we will even provide improved bounds for passive testing in the process as well.

2.1.2 Our Results

We have two types of results. Our first results, on the testability of unions of intervals and linear threshold functions, show that it is indeed possible to test properties of interest to the learning community efficiently in the active model. Our next results, concerning the testing of disjoint unions of properties and a new notion of testing dimension, examine the active testing model from a more abstract point of view. We describe these results and some of their applications below.

Testing Unions of Intervals. The function $f : [0,1] \to \{0,1\}$ is a *union of d intervals* if there are at most d non-overlapping intervals $(\ell_1, u_1), \ldots, (\ell_d, u_d)$ such that $f(x) = 1$ iff $\ell_i \le x \le u_i$ for some $i \in [d]$. The VC dimension of this class is $2d$, so learning a union of d intervals requires at least $\Omega(d)$ queries. By contrast, we show that testing unions of d intervals can be done with a number of label requests that is *independent* of d, for any distribution D:

Theorem 2.4. *Testing unions of d intervals in the active testing model can be done using only $O(1/\epsilon^3)$ queries. In the case of the uniform distribution, we further need only $O(\sqrt{d}/\epsilon^5)$ unlabeled examples.*

We note that Theorem 2.4 not only gives the first result for testing unions of intervals in the active testing model, but it also improves on the previous best results for testing this class in the membership query and passive models. Previous testers used $O(1)$ queries in the membership query model and $O(\sqrt{d})$ samples in the passive model, but applied only to a relaxed setting in which only functions that were ϵ far from unions of $d' = d/\epsilon$ intervals had to be rejected

with high probability [Kearns and Ron, 2000]. Our tester immediately yields the same query bound as a function of d (active testing with $O(\sqrt{d})$ unlabeled examples directly implies passive testing with $O(\sqrt{d})$ labeled examples) but rejects any function that is ϵ-far from unions of $d' = d$ intervals. Note also that Kearns and Ron [Kearns and Ron, 2000] show that $\Omega(\sqrt{d})$ samples are required to test unions of d intervals in the passive model, and so our bound on the number of unlabeled examples in Theorem 2.4 is optimal in terms of d.

The proof of Theorem 2.4 relies on a new *noise sensitivity* characterization of the class of unions of d intervals. That is, we show that all unions of d intervals have low noise sensitivity while all functions that are far from this class have noticeably larger noise sensitivity and introduce a tester that estimates the noise sensitivity of the input function. We describe these results in Section 2.2.

Testing Linear Threshold Functions. We next study the problem of testing linear threshold functions (or LTFs), namely the class of boolean functions $f : R^n \to \{0, 1\}$ of the form $f(x) = \text{sgn}(w_1 x_1 + \cdots + w_n x_n - \theta)$ where $w_1, \ldots, w_n, \theta \in \mathbb{R}$. LTFs can be tested with $O(1)$ queries in the membership query model [Matulef, O'Donnell, Rubinfeld, and Servedio, 2009]. While we show this is not possible in the active testing model, we nonetheless show we can substantially improve over the number of label requests needed for *learning*. In particular, learning LTFs requires $\Theta(n)$ labeled examples, even over the Gaussian distribution [Long, 1995]. We show that the query and sample complexity for *testing* LTFs is significantly better:

Theorem 2.5. *We can efficiently test LTFs under the Gaussian distribution with $\tilde{O}(\sqrt{n})$ labeled examples in both active and passive testing models. Furthermore, we have lower bounds of $\tilde{\Omega}(n^{1/3})$ and $\tilde{\Omega}(\sqrt{n})$ on the number of labels needed for active and passive testing respectively.*

The proof of the upper bound in the theorem relies on a recent characterization of LTFs by the Hermite weight distribution of the function [Matulef, O'Donnell, Rubinfeld, and Servedio, 2009] as well as a new concentration of measure result for U-statistics. The proof of the lower bound involves analyzing the distance between the label distribution of an LTF formed by a Gaussian

17

weight vector and the label distribution of a random noise function. See Section 2.3 for details.

Testing Disjoint Unions of Testable Properties. Given a collection of properties $\mathcal{P}_i$, a natural way to combine them is via their disjoint union. E.g., perhaps our data falls into N well-separated regions, and while we suspect our data overall may not be linearly separable, we believe it may be linearly separable (by a different separator) in each region. We show that if each individual property $\mathcal{P}_i$ is testable (in this case, $\mathcal{P}_i$ is the LTF property) then their disjoint union $\mathcal{P}$ is testable as well, with only a very small increase in the total number of queries. It is worth noting that this property does *not* hold for passive testing. We present this result in Section 2.4, and use it inside our testers for semi-supervised learning properties discussed below.

Testing Semi-Supervised Learning Assumptions. Two common assumptions considered in semi-supervised learning [Chapelle, Schlkopf, and Zien, 2006] and active learning [Dasgupta, 2011] are (a) if data happens to cluster then points in the same cluster should have the same label, and (b) there should be some large margin γ of separation between the positive and negative region (but without assuming the target is necessarily a linear threshold function). Here, we show that for both properties, active testing can be done with $O(1)$ label requests, even though these classes contain functions of high complexity so learning (even semi-supervised or active) requires substantially more labeled examples. Our results for the margin assumption use the cluster tester as a subroutine, along with analysis of an appropriate weighted graph defined over the data. We present our results in Section 2.4 but for space reasons, defer analysis to Appendix 2.11.

General Testing Dimensions. We develop a general notion of the *testing dimension* of a given property with respect to a given distribution. We do this for both passive and active testing models. We show these dimensions characterize (up to constant factors) the intrinsic number of label requests needed to test the given property with respect to the given distribution in the corresponding model. For the case of active testing we also provide a simpler notion that characterizes

18

whether testing with $O(1)$ label requests is possible. We present the dimension definitions and analysis in Section 2.5.

The lower bounds in this paper are given by proving lower bounds on these dimension quantities. In Section 2.5.1, we prove (as mentioned above) that for the class of dictator functions, active testing cannot be done with fewer queries than the number of examples needed for learning, even for the problem of distinguishing dictator functions from truly random functions. This result additionally implies that any class that contains dictator functions (and is not so large as to contain almost all functions) requires $\Omega(\log n)$ queries to test in the active model, including decision trees, functions of low Fourier degree, juntas, DNFs, etc. In Section 2.5.2, we complete the proofs of the lower bounds in Theorem 2.5 on the number of queries required to test linear threshold functions.

2.2 Testing Unions of Intervals

In this section, we prove Theorem 2.4 that we can test unions of d intervals in the active testing model using only $O(1/\epsilon^3)$ label requests, and furthermore, over the uniform distribution, using only $O(\sqrt{d}/\epsilon^5)$ unlabeled samples. We begin with the case that the underlying distribution is uniform over $[0, 1]$, and afterwards show how to generalize to arbitrary distributions. Our tester exploits the fact that unions of intervals have a *noise sensitivity* characterization.

Definition 2.6. *Fix $\delta > 0$. The* local δ-noise sensitivity *of the function $f : [0, 1] \to \{0, 1\}$ at $x \in [0, 1]$ is* $\mathrm{NS}_\delta(f, x) = \mathrm{Pr}_{y \sim_\delta x}[f(x) \neq f(y)]$, *where $y \sim_\delta x$ represents a draw of y uniform in $(x - \delta, x + \delta) \cap [0, 1]$. The* noise sensitivity *of f is*

$$\mathbb{NS}_\delta(f) = \Pr_{x, y \sim_\delta x}[f(x) \neq f(y)]$$

or, equivalently, $\mathbb{NS}_\delta(f) = \mathbb{E}_x \mathrm{NS}_\delta(f, x)$.

A simple argument shows that unions of d intervals have (relatively) low noise sensitivity:

Proposition 2.7. *Fix $\delta > 0$ and let $f : [0, 1] \to \{0, 1\}$ be a union of d intervals. Then $\mathbb{NS}_\delta(f) \le d\delta$.*

Proof sketch. Draw $x \in [0, 1]$ uniformly at random and $y \sim_\delta x$. The inequality $f(x) \ne f(y)$ can only hold when a boundary $b \in [0, 1]$ of one of the d intervals in f lies in between x and y. For any point $b \in [0, 1]$, the probability that $x < b < y$ or $y < b < x$ is at most $\frac{\delta}{2}$, and there are at most $2d$ boundaries of intervals in f, so the proposition follows from the union bound. □

Interestingly, the converse of the proposition statement is approximately true: for δ small enough, every function that has noise sensitivity not much larger than $d\delta$ is close to being a union of d intervals. (Full proof in Appendix 2.7).

Lemma 2.8. *Fix $\delta = \frac{\epsilon^2}{32d}$. Let $f : [0, 1] \to \{0, 1\}$ be a function with noise sensitivity bounded by $\mathbb{NS}_\delta(f) \le d\delta(1 + \frac{\epsilon}{4})$. Then f is ϵ-close to a union of d intervals.*

Proof outline. The proof proceeds in two steps. First, we construct a function $g : [0, 1] \to \{0, 1\}$ that is $\frac{\epsilon}{2}$-close to f and is a union of at most $d(1 + \frac{\epsilon}{4})$ intervals. We then show that g – and every other function that is a union of at most $d(1 + \frac{\epsilon}{4})$ intervals – is $\frac{\epsilon}{2}$-close to a union of d intervals.

To construct the function g, we consider the "smoothed" function $f_\delta : [0, 1] \to [0, 1]$ obtained by taking the convolution of f and a uniform kernel of width 2δ. We define τ to be some appropriately small parameter. When $f_\delta(x) \le \tau$, then this means that nearly all the points in the δ-neighborhood of x have the value 0 in f, so we set $g(x) = 0$. Similarly, when $f_\delta(x) \ge 1 - \tau$, then we set $g(x) = 1$. (This procedure removes any "local noise" that might be present in f.) This leaves all the points x where $\tau < f_\delta(x) < 1 - \tau$. Let us call these points *undefined*. For each such point x we take the largest value $y \le x$ that is defined and set $g(x) = g(y)$.

The key technical part of the proof involves showing that the construction described above yields a function g that is ϵ-close to f and that is a union of $d(1 + \frac{\epsilon}{4})$ intervals. This is done with standard tools from function analysis and probability theory. Due to space constraints, we defer the details to Appendix 2.7. □

The noise sensitivity characterization of unions of intervals obtained by Proposition 2.7 and Lemma 2.8 suggest a natural approach for building a tester: design an algorithm that estimates the noise sensitivity of the input function and accepts iff this noise sensitivity is small enough. This is indeed what we do:

UNION OF INTERVALS TESTER(f, d, ϵ)

Parameters: $\delta = \frac{\epsilon^2}{32d}, r = O(\epsilon^{-3})$.

1. For rounds $i = 1, \ldots, r$,

    1.1 Draw $x \in [0, 1]$ uniformly at random.

    1.2 Draw samples until we obtain $y \in (x - \delta, x + \delta)$.

    1.3 Set $Z_i = \mathbf{1}[f(x) \neq f(y)]$.

2. **Accept** iff $\frac{1}{r} \sum Z_i \leq d\delta(1 + \frac{\epsilon}{8})$.

The algorithm makes $2r = O(\epsilon^{-3})$ queries to the function. Since a draw in Step 1.2 is in the desired range with probability 2δ, the number of samples drawn by the algorithm is a random variable with very tight concentration around $r(1 + \frac{1}{2\delta}) = O(d/\epsilon^5)$. The draw in Step 1.2 also corresponds to choosing $y \sim_\delta x$. As a result, the probability that $f(x) \neq f(y)$ in a given round is exactly $\mathbb{NS}_\delta(f)$, and the average $\frac{1}{r} \sum Z_i$ is an unbiased estimate of the noise sensitivity of f. By Proposition 2.7, Lemma 2.8, and Chernoff bounds, the algorithm therefore errs with probability less than $\frac{1}{3}$ provided that $r > c \cdot 1/d\delta\epsilon = c \cdot 32/\epsilon^3$ for some suitably large constant c.

Improved unlabeled sample complexity: Notice that by changing Steps 1.1-1.2 slightly to pick the first pair (x, y) such that $|x - y| < \delta$, we immediately improve the unlabeled sample complexity to $O(\sqrt{d}/\epsilon^5)$ without affecting the analysis. In particular, this procedure is equivalent to picking $x \in [0, 1]$ then $y \sim_\delta x$.[4] As a result, up to $poly(1/\epsilon)$ terms, we also improve over the *passive testing* bounds of Kearns and Ron [Kearns and Ron, 2000] which are able only to distinguish the case that f is a union of d intervals from the case that f is ϵ-far from being a

    [4]Except for events of $O(\delta)$ probability mass at the boundary.

union of d/ϵ intervals. (Their results use $O(\sqrt{d}/\epsilon^{1.5})$ examples.) Kearns and Ron [Kearns and Ron, 2000] show that $\Omega(\sqrt{d})$ examples are necessary for passive testing, so in terms of d this is optimal.

Active Tester Over Arbitrary Distributions: We can reduce the problem of testing over general distributions to that of testing over the uniform distribution on $[0, 1]$ by using the CDF of the distribution D. In particular, given point x, define $p_x = \mathrm{Pr}_{y \sim D}[y \leq x]$. So, for x drawn from D, p_x is uniform in $[0, 1]$.[5] As a result we can just replace Step 1.2 in the tester with sampling until we obtain y such that $p_y \in (p_x - \delta, p_x + \delta)$. The only issue is that we do not know the p_x and p_y values exactly. However, VC-dimension bounds for initial intervals on the line imply that if we sample $O(\epsilon^{-6}\delta^{-2})$ unlabeled examples, with high probability the estimates $\hat{p}_x$ computed with respect to the sample (the fraction of points in the *sample* that are $\leq x$) will be within $O(\epsilon^3\delta)$ of the correct p_x values for all points x. This in turn implies that the noise-sensitivity estimates are sufficiently accurate that the procedure works as before.

Putting these results together, we have Theorem 2.4.

2.3   Testing Linear Threshold Functions

In the last section, we saw how unions of intervals are characterized by a statistic of the function – namely, its noise sensitivity – that can be estimated with few queries and used this to build our tester. In this section, we follow the same high-level approach for testing linear threshold functions. In this case, however, the statistic we will estimate is not noise sensitivity but rather the sum of squares of the degree-1 Hermite coefficients of the function.

Definition 2.9. *The* Hermite polynomials *are a set of polynomials* $h_0(x) = 1, h_1(x) = x, h_2(x) = \frac{1}{\sqrt{2}}(x^2 - 1), \ldots$ *that form a complete orthogonal basis for (square-integrable) functions* $f : \mathbb{R} \to \mathbb{R}$ *over the inner product space defined by the inner product* $\langle f, g \rangle = \mathbb{E}_x[f(x)g(x)]$, *where*

---

[5]We are assuming here that D is continuous and has a pdf. If D has point masses, then instead define $p_x^L = \mathrm{Pr}_y[y < x]$ and $p_x^U = \mathrm{Pr}_y[y \leq x]$ and select p_x uniformly in $[p_x^L, p_x^U]$.

the expectation is over the standard Gaussian distribution $\mathcal{N}(0,1)$. For any $S \in \mathbb{N}^n$, define $H_S = \prod_{i=1}^n h_{S_i}(x_i)$. The Hermite coefficient *of $f : \mathbb{R}^n \to \mathbb{R}$ corresponding to S is $\hat{f}(S) = \langle f, H_S \rangle = \mathbb{E}_x[f(x)H_S(x)]$ and the* Hermite decomposition *of f is $f(x) = \sum_{S \in \mathbb{N}^n} \hat{f}(S)H_S(x)$. The* degree *of the coefficient $\hat{f}(S)$ is $|S| := \sum_{i=1}^n S_i$.*

The connection between linear threshold functions and the Hermite decomposition of functions is revealed by the following key lemma of Matulef et al. [Matulef, O'Donnell, Rubinfeld, and Servedio, 2009].

Lemma 2.10 (Matulef et al. [Matulef, O'Donnell, Rubinfeld, and Servedio, 2009])**.** *There is an explicit continuous function $W : \mathbb{R} \to \mathbb{R}$ with bounded derivative $\|W'\|_\infty \leq 1$ and peak value $W(0) = \frac{2}{\pi}$ such that every linear threshold function $f : \mathbb{R}^n \to \{-1,1\}$ satisfies $\sum_{i=1}^n \hat{f}(e_i)^2 = W(\mathbb{E}_x f)$. Moreover, every function $g : \mathbb{R}^n \to \{-1,1\}$ that satisfies $|\sum_{i=1}^n \hat{g}(e_i)^2 - W(\mathbb{E}_x g)| \leq 4\epsilon^3$, is ϵ-close to being a linear threshold function.*

In other words, Lemma 2.10 shows that $\sum_i \hat{f}(e_i)^2$ characterizes linear threshold functions. To test LTFs, it suffices to estimate this value (and the expected value of the function) with enough accuracy. Matulef et al. [Matulef, O'Donnell, Rubinfeld, and Servedio, 2009] showed that $\sum_i \hat{f}(e_i)^2$ can be estimated with a number of queries that is independent of n by querying f on pairs $x, y \in \mathbb{R}^n$ where the marginal distributions on x and y are both the standard Gaussian distribution and where $\langle x, y \rangle = \eta$ for some small (but constant) $\eta > 0$. Unfortunately, the same approach does not work in the active testing model since with high probability, all pairs of samples that we can query have inner product $|\langle x, y \rangle| \leq O(\frac{1}{\sqrt{n}})$. Instead, we rely on the following result.

Lemma 2.11. *For any function $f : \mathbb{R}^n \to \mathbb{R}$, we have $\sum_{i=1}^n \hat{f}(e_i)^2 = \mathbb{E}_{x,y}[f(x)f(y)\langle x, y \rangle]$ where $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$ is the standard vector dot product.*

Proof. Applying the Hermite decomposition of f and linearity of expectation,

$$\mathbb{E}_{x,y}[f(x)f(y)\langle x, y \rangle] = \sum_{i=1}^n \sum_{S,T \in \mathbb{N}^n} \hat{f}(S)\hat{f}(T)\mathbb{E}_x[H_S(x)x_i]\mathbb{E}_y[H_T(y)y_i].$$

23

By definition, $x_i = h_1(x_i) = H_{e_i}(x)$. The orthonormality of the Hermite polynomials therefore guarantees that $\mathbb{E}_x[H_S(x)H_{e_i}(x)] = \mathbf{1}[S\!=\!e_i]$. Similarly, $\mathbb{E}_y[H_T(y)y_i] = \mathbf{1}[T\!=\!e_i]$. $\qquad\square$

A natural idea for completing our LTF tester is to simply sample pairs $x, y \in \mathbb{R}^n$ independently at random and evaluating $f(x)f(y)\langle x, y\rangle$ on each pair. While this approach does give an unbiased estimate of $\mathbb{E}_{x,y}[f(x)f(y)\langle x, y\rangle]$, it has poor query efficiency: To get enough accuracy, we need to repeat this sampling strategy $\Omega(n)$ times. (That is, the query complexity of this sampling approach is the same as that of *learning* LTFs.)

We can improve the query complexity of the sampling strategy by instead using *U-statistics*. The U-statistic (of order 2) with symmetric kernel function $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is

$$U_g^m(x^1, \ldots, x^m) := \binom{m}{2}^{-1} \sum_{1 \leq i < j \leq m} g(x^i, x^j).$$

Tight concentration bounds are known for U-statistics with well-behaved kernel functions. In particular, by setting $g(x, y) = f(x)f(y)\langle x, y\rangle \mathbf{1}[|\langle x, y\rangle| < \tau]$ to be an appropriately truncated kernel for estimating $\mathbb{E}[f(x)f(y)\langle x, y\rangle]$, we can apply a Bernstein-type inequality due to Arcones [Arcones, 1995] to show that $O(\sqrt{n})$ samples are sufficient to estimate $\sum_i \hat{f}(e_i)^2$ with sufficient accuracy. As a result, the following algorithm is a valid tester for LTFs.

LTF TESTER(f, ϵ)

Parameters: $\tau = \sqrt{4n \log(4n/\epsilon^3)}$, $m = 800\tau/\epsilon^3 + 32/\epsilon^6$.

1. Draw $x^1, x^2, \ldots, x^m$ independently at random from $\mathbb{R}^n$.

2. Query $f(x^1), f(x^2), \ldots, f(x^m)$.

3. Set $\tilde{\mu} = \frac{1}{m}\sum_{i=1}^m f(x^i)$.

4. Set $\tilde{\nu} = \binom{m}{2}^{-1} \sum_{i \neq j} f(x^i)f(x^j)\langle x^i, x^j\rangle \cdot \mathbf{1}[|\langle x^i, x^j\rangle| \leq \tau]$.

5. **Accept** iff $|\tilde{\nu} - W(\tilde{\mu})| \leq 2\epsilon^3$.

The algorithm queries the function only on inputs that are all independently drawn at random from the n-dimensional Gaussian distribution. As a result, this tester works in both the active

and passive testing models. For the complete proof of the correctness of the algorithm, see Appendix 2.8.

2.4 Testing Disjoint Unions of Testable Properties

We now show that active testing has the feature that a disjoint union of testable properties is testable, with a number of queries that is independent of the size of the union; this feature does not hold for passive testing. In addition to providing insight into the distinction between the two models, this fact will be useful in our analysis of semi-supervised learning-based properties mentioned below and discussed more fully in Appendix 2.11.

Specifically, given properties $\mathcal{P}_1, \ldots, \mathcal{P}_N$ over domains $X_1, \ldots, X_N$, define their disjoint union $\mathcal{P}$ over domain $X = \{(i, x) : i \in [N], x \in X_i\}$ to be the set of functions f such that $f(i, x) = f_i(x)$ for some $f_i \in \mathcal{P}_i$. In addition, for any distribution D over X, define D_i to be the conditional distribution over X_i when the first component is i. If each $\mathcal{P}_i$ is testable over D_i then $\mathcal{P}$ is testable over D with only small overhead in the number of queries:

Theorem 2.12. *Given properties $\mathcal{P}_1, \ldots, \mathcal{P}_N$, if each $\mathcal{P}_i$ is testable over D_i with $q(\epsilon)$ queries and $U(\epsilon)$ unlabeled samples, then their disjoint union $\mathcal{P}$ is testable over the combined distribution D with $O(q(\epsilon/2) \cdot (\log^3 \frac{1}{\epsilon}))$ queries and $O(U(\epsilon/2) \cdot (\frac{N}{\epsilon} \log^3 \frac{1}{\epsilon}))$ unlabeled samples.*

Proof. See Appendix 2.9. □

As a simple example, consider $\mathcal{P}_i$ to contain just the constant functions **1** and **0**. In this case, $\mathcal{P}$ is equivalent to what is often called the "cluster assumption," used in semi-supervised and active learning [Chapelle, Schlkopf, and Zien, 2006, Dasgupta, 2011], that if data lies in some number of clearly identifiable clusters, then all points in the same cluster should have the same label. Here, each $\mathcal{P}_i$ individually is easily testable (even passively) with $O(1/\epsilon)$ labeled samples, so Theorem 2.12 implies the cluster assumption is testable with $poly(1/\epsilon)$ queries.[6] However, it

---

[6]Since the $\mathcal{P}_i$ are so simple in this case, one can actually test with only $O(1/\epsilon)$ queries.

is not hard to see that passive testing with $poly(1/\epsilon)$ samples is not possible and in fact requires $\Omega(\sqrt{N}/\epsilon)$ labeled examples.[7]

We build on this to produce testers for other properties often used in semi-supervised learning. In particular, we prove the following result about testing the margin property (See Appendix 2.11 for definitions and analysis).

Theorem 2.13. *For any γ, $\gamma' = \gamma(1 - 1/c)$ for constant $c > 1$, for data in the unit ball in R^d for constant d, we can distinguish the case that D_f has margin γ from the case that D_f is ϵ-far from margin γ' using Active Testing with $O(1/(\gamma^{2d}\epsilon^2))$ unlabeled examples and $O(1/\epsilon)$ label requests.*

2.5 General Testing Dimensions

The previous sections have discussed upper and lower bounds for a variety of classes. Here, we define notions of *testing dimension* for passive and active testing that characterize (up to constant factors) the number of labels needed for testing to succeed, in the corresponding testing protocols. These will be distribution-specific notions (like SQ dimension in learning), so let us fix some distribution D over the instance space X, and furthermore fix some value ϵ defining our goal. I.e., our goal is to distinguish the case that $dist_D(f, \mathcal{P}) = 0$ from the case $dist_D(f, \mathcal{P}) \geq \epsilon$.

For a given set S of unlabeled points, and a distribution π over boolean functions, define π_S to be the distribution over labelings of S induced by π. That is, for $y \in \{0,1\}^{|S|}$ let $\pi_S(y) = \Pr_{f \sim \pi}[f(S) = y]$. We now use this to define a distance between distributions. Specifically, given a set of unlabeled points S and two distributions π and π' over boolean functions, define

$$D_S(\pi, \pi') = (1/2) \sum_{y \in \{0,1\}^{|S|}} |\pi_S(y) - \pi'_S(y)|,$$

[7]Specifically, suppose region 1 has $1 - 2\epsilon$ probability mass with $f_1 \in \mathcal{P}_1$, and suppose the other regions equally share the remaining 2ϵ probability mass and either (a) are each pure but random (so $f \in \mathcal{P}$) or (b) are each 50/50 (so f is ϵ-far from $\mathcal{P}$). Distinguishing these cases requires seeing at least two points with the same index $i \neq 1$, yielding the $\Omega(\sqrt{N}/\epsilon)$ bound.

to be the variation distance between π and π' induced by S. Finally, let Π_0 be the set of all distributions π over functions in $\mathcal{P}$, and let set Π_ϵ be the set of all distributions π' in which a $1 - o(1)$ probability mass is over functions at least ϵ-far from $\mathcal{P}$. We are now ready to formulate our notions of dimension.

Definition 2.14. *Define the passive testing dimension, $d_{passive}$, as the largest $q \in \mathbb{N}$ such that,*

$$\sup_{\pi \in \Pi_0} \sup_{\pi' \in \Pi_\epsilon} \Pr_{S \sim D^q} (\mathrm{D}_S(\pi, \pi') > 1/4) \leq 1/4.$$

That is, there exist distributions π and π' such that a random set S of $d_{passive}$ examples has a reasonable probability (at least $3/4$) of having the property that one cannot reliably distinguish a random function from π versus a random function from π' from just the labels of S. From the definition it is fairly immediate that $\Omega(d_{passive})$ examples are *necessary* for passive testing; in fact, $O(d_{passive})$ are sufficient as well.

Theorem 2.15. *The sample complexity of passive testing is $\Theta(d_{passive})$.*

Proof. See Appendix 2.10. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

For the case of active testing, there are two complications. First, the algorithms can examine their entire $poly(n)$-sized unlabeled sample before deciding which points to query, and secondly they may in principle determine the next query based on the responses to the previous ones (even though all our algorithmic results do not require this feature). If we merely want to distinguish those properties that are actively testable with $O(1)$ queries from those that are not, then the second complication disappears and the first is simplified as well, and the following coarse notion of dimension suffices.

Definition 2.16. *Define the coarse active testing dimension, d_{coarse}, as the largest $q \in \mathbb{N}$ such that,*

$$\sup_{\pi \in \Pi_0} \sup_{\pi' \in \Pi_\epsilon} \Pr_{S \sim D^q} (\mathrm{D}_S(\pi, \pi') > 1/4) \leq 1/n^q.$$

Theorem 2.17. *If $d_{coarse} = O(1)$ the active testing of $\mathcal{P}$ can be done with $O(1)$ queries, and if $d_{coarse} = \omega(1)$ then it cannot.*

Proof. See Appendix 2.10. □

To achieve a more fine-grained characterization of active testing we consider a slightly more involved quantity, as follows. First, recall that given an unlabeled sample U and distribution π over functions, we define π_U as the induced distribution over labelings of U. We can view this as a distribution over *unlabeled* examples in $\{0, 1\}^{|U|}$. Now, given two distributions over functions π, π', define $\text{Fair}(\pi, \pi', U)$ to be the distribution over *labeled* examples (y, ℓ) defined as: with probability $1/2$ choose $y \sim \pi_U$, $\ell = 1$ and with probability $1/2$ choose $y \sim \pi'_U$, $\ell = 0$. Thus, for a given unlabeled sample U, the sets Π_0 and Π_ϵ define a *class* of fair distributions over labeled examples. The active testing dimension, roughly, asks how well this class can be approximated by the class of low-depth decision trees. Specifically, let DT_k denote the class of decision trees of depth at most k. The active testing dimension for a given number u of allowed unlabeled examples is as follows:

Definition 2.18. *Given a number $u = poly(n)$ of allowed unlabeled examples, we define the active testing dimension, $d_{active}(u)$, as the largest $q \in \mathbb{N}$ such that*

$$\sup_{\pi \in \Pi_0} \sup_{\pi' \in \Pi_\epsilon} \Pr_{U \sim D^u} \left(\text{err}^*(\text{DT}_q, \text{Fair}(\pi, \pi', U)) < 1/4 \right) \leq 1/4,$$

where $\text{err}^*(H, P)$ is the error of the optimal function in H with respect to data drawn from distribution P over labeled examples.

Theorem 2.19. *Active testing with failure probability $\frac{1}{8}$ using u unlabeled examples requires $\Omega(d_{active}(u))$ label queries, and furthermore can be done with $O(u)$ unlabeled examples and $O(d_{active}(u))$ label queries.*

Proof. See Appendix 2.10. □

We now use these notions of dimension to prove lower bounds for testing several properties.

28

2.5.1 Application: Dictator functions

We now prove Theorem 2.3 that active testing of dictatorships over the uniform distribution requires $\Omega(\log n)$ queries by proving a $\Omega(\log n)$ lower bound on $d_{active}(u)$ for any $u = poly(n)$; in fact, this result holds even for the specific choice of π' as random noise (the uniform distribution over all functions).

Proof of Theorem 2.3. Define π and π' to be uniform distributions over the dictator functions and over all boolean functions, respectively. In particular, π is the distribution obtained by choosing $i \in [n]$ uniformly at random and returning the function $f : \{0,1\}^n \to \{0,1\}$ defined by $f(x) = x_i$. Fix S to be a set of q vectors in $\{0,1\}^n$. This set can be viewed as a $q \times n$ boolean-valued matrix. We write $c_1(S), \ldots, c_n(S)$ to represent the columns of this matrix. For any $y \in \{0,1\}^q$,

$$\pi_S(y) = \frac{|\{i \in [n] : c_i(S) = y\}|}{n} \qquad \text{and} \qquad \pi'_S(y) = 2^{-q}.$$

By Lemma 2.21, to prove that $d_{active} \geq \frac{1}{2} \log n$, it suffices to show that when $q < \frac{1}{2} \log n$ and U is a set of n^c vectors chosen uniformly and independently at random from $\{0,1\}^n$, then with probability at least $\frac{3}{4}$, every set $S \subseteq U$ of size $|S| = q$ and every $y \in \{0,1\}^q$ satisfy $\pi_S(y) \leq \frac{6}{5} 2^{-q}$. (This is like a stronger version of d_{coarse} where $D_S(\pi, \pi')$ is replaced with an L_∞ distance.)

Consider a set S of q vectors chosen uniformly and independently at random from $\{0,1\}^n$. For any vector $y \in \{0,1\}^q$, the expected number of columns of S that are equal to y is $n2^{-q}$. Since the columns are drawn independently at random, Chernoff bounds imply that

$$\Pr\left[\pi_S(y) > \tfrac{6}{5} 2^{-q}\right] \leq e^{-(\frac{1}{5})^2 n 2^{-q}/3} < e^{-\frac{1}{75} n 2^{-q}}.$$

By the union bound, the probability that there exists a vector $y \in \{0,1\}^q$ such that more than $\frac{6}{5} n 2^{-q}$ columns of S are equal to y is at most $2^q e^{-\frac{1}{75} n 2^{-q}}$. Furthermore, when U is defined as above, we can apply the union bound once again over all subsets $S \subseteq U$ of size $|S| = q$ to obtain $\Pr[\exists S, y : \pi_S(y) > \frac{6}{5} 2^{-q}] < n^{cq} \cdot 2^q \cdot e^{-\frac{1}{75} n 2^{-q}}$. When $q \leq \frac{1}{2} \log n$, this probability is bounded

above by $e^{\frac{c}{2}\log^2 n + \frac{1}{2}\log n - \frac{1}{75}\sqrt{n}}$, which is less than $\frac{1}{4}$ when n is large enough, as we wanted to show. $\qquad\square$

2.5.2 Application: LTFs

The testing dimension also lets us prove the lower bounds in Theorem 2.5 regarding the query complexity for testing linear threshold functions. Specifically, those bounds follow directly from the following result.

Theorem 2.20. *For linear threshold functions under the standard n-dimensional Gaussian distribution, $d_{passive} = \Omega(\sqrt{n/\log(n)})$ and $d_{active} = \Omega((n/\log(n))^{1/3})$.*

Let us give a brief overview of the strategies used to obtain the $d_{passive}$ and d_{active} bounds. The complete proofs for both results, as well as a simpler proof that $d_{coarse} = \Omega((n/\log n)^{1/3})$, can be found in Appendix 2.10.4.

For both results, we set π to be a distribution over LTFs obtained by choosing $w \sim \mathcal{N}(0, I_{n\times n})$ and outputting $f(x) = \mathrm{sgn}(w \cdot x)$. Set π' to be the uniform distribution over all functions—i.e., for any $x \in \mathbb{R}^n$, the value of $f(x)$ is uniformly drawn from $\{0, 1\}$ and is independent of the value of f on other inputs.

To bound $d_{passive}$, we bound the total variation distance between the distribution of $Xw/\sqrt{n}$ given X, and the standard normal $\mathcal{N}(0, I_{n\times n})$. If this distance is small, then so must be the distance between the distribution of $\mathrm{sgn}(Xw)$ and the uniform distribution over label sequences.

Our strategy for bounding d_{active} is very similar to the one we used to prove the lower bound on the query complexity for testing dictator functions in the last section. Again, we want to apply Lemma 2.21. Specifically, we want to show that when $q \le o((n/\log(n))^{1/3})$ and U is a set of n^c vectors drawn independently from the n-dimensional standard Gaussian distribution, then with probability at least $\frac{3}{4}$, every set $S \subseteq U$ of size $|S| = q$ and almost all $x \in \mathbb{R}^q$, we have $\pi_S(x) \le \frac{6}{5}2^{-q}$. The difference between this case and the lower bound for dictator functions is that we now rely on strong concentration bounds on the spectrum of random matrices [Vershynin,

30

2012] to obtain the desired inequality.

2.6   Proof of a Property Testing Lemma

The following lemma is a generalization of a lemma that is widely used for proving lower bounds
in property testing [Fischer, 2001, Lem. 8.3]. We use this lemma to prove the lower bounds on
the query complexity for testing dictator functions and testing linear threshold functions.

Lemma 2.21. *Let π and π' be two distributions on functions $X \to \mathbb{R}$. Fix $U \subseteq X$ to be a set
of allowable queries. Suppose that for any $S \subseteq U$, $|S| = q$, there is a set $E_S \subseteq \mathbb{R}^q$ (possibly
empty) satisfying $\pi_S(E_S) \leq \frac{1}{5}2^{-q}$ such that*

$$\pi_S(y) < \tfrac{6}{5}\pi'_S(y) \ \text{for every } y \in \mathbb{R}^q \setminus E_S.$$

Then $\operatorname{err}^*(\mathrm{DT}_q, \operatorname{Fair}(\pi, \pi', U)) > 1/4$.

Proof. Consider any decision tree $\mathcal{A}$ of depth q. Each internal node of the tree consists of a
query $y \in U$ and a subset $T \subseteq \mathbb{R}$ such that its children are labeled by T and $\mathbb{R} \setminus T$, respectively.
The leaves of the tree are labeled with either "accept" or "reject", and let L be the set of leaves
labeled as accept. Each leaf $\ell \in L$ corresponds to a set $S_\ell \subseteq U^q$ of queries and a subset $T_\ell \subseteq \mathbb{R}^\ell$,
where $f : X \to \mathbb{R}$ leads to the leaf ℓ iff $f(S_\ell) \in T_\ell$. The probability that $\mathcal{A}$ (correctly) accepts
an input drawn from π is

$$a_1 = \sum_{\ell \in L} \int_{T_\ell} \pi_{S_\ell}(y) dy.$$

Similarly, the probability that $\mathcal{A}$ (incorrectly) accepts an input drawn from π' is

$$a_2 = \sum_{\ell \in L} \int_{T_\ell} \pi'_{S_\ell}(y) dy.$$

The difference between the two rejection probabilities is bounded above by

$$a_1 - a_2 \leq \sum_{\ell \in L} \int_{T_\ell \setminus E_{S_\ell}} \pi_{S_\ell}(y) - \pi'_{S_\ell}(y) dy + \sum_{\ell \in L} \int_{T_\ell \cap E_{S_\ell}} \pi_{S_\ell}(y) dy.$$

31

The conditions in the statement of the lemma then imply that

$$a_1 - a_2 < \sum_{\ell \in L} \int_{T_\ell} \tfrac{1}{6} \pi_{S_\ell}(y) dy + \tfrac{5}{6} \sum_{\ell} \int_{E_{S_\ell}} \pi_{S_\ell}(y) dy \leq \tfrac{1}{3}.$$

To complete the proof, we note that $\mathcal{A}$ errs on an input drawn from $\mathrm{Fair}(\pi, \pi', U)$ with probability

$$\tfrac{1}{2}(1 - a_1) + \tfrac{1}{2}a_2 = \tfrac{1}{2} - \tfrac{1}{2}(a_1 - a_2) > \tfrac{1}{3}. \qquad \square$$

2.7 Proofs for Testing Unions of Intervals

In this section we complete the proofs of the technical results in Section 2.2.

Proposition 2.7 (Restated)**.** *Fix $\delta > 0$ and let $f : [0, 1] \to \{0, 1\}$ be a union of d intervals. Then* $\mathbb{NS}_\delta(f) \leq d\delta.$

Proof. For any fixed $b \in [0, 1]$, the probability that $x < b < y$ when $x \sim U(0, 1)$ and $y \sim U(x - \delta, x + \delta)$ is

$$\Pr_{x,y}[x < b < y] = \int_0^\delta \Pr_{y \sim U(b-t-\delta, b-t+\delta)}[y \geq b] dt = \int_0^\delta \frac{\delta - t}{2\delta} dt = \frac{\delta}{4}.$$

Similarly, $\Pr_{x,y}[y < b < x] = \frac{\delta}{4}$. So the probability that b lies between x and y is at most $\frac{\delta}{2}$.

When f is the union of d intervals, $f(x) \neq f(y)$ only if at least one of the boundaries $b_1, \ldots, b_{2d}$ of the intervals of f lies in between x and y. So by the union bound, $\Pr[f(x) \neq f(y)] \leq 2d(\delta/2) = d\delta$. Note that if b is within distance δ of 0 or 1, the probability is only lower. $\qquad \square$

Lemma 2.8 (Restated)**.** *Fix $\delta = \frac{\epsilon^2}{32d}$. Let $f : [0, 1] \to \{0, 1\}$ be any function with noise sensitivity* $\mathbb{NS}_\delta(f) \leq d\delta(1 + \frac{\epsilon}{4})$. *Then f is ϵ-close to a union of d intervals.*

Proof. The proof proceeds in two steps: We first show that f is $\frac{\epsilon}{2}$-close to a union of $d(1 + \frac{\epsilon}{2})$ intervals, then we show that every union of $d(1 + \frac{\epsilon}{2})$ intervals is $\frac{\epsilon}{2}$-close to a union of d intervals.

32

Consider the "smoothed" function $f_\delta : [0, 1] \to [0, 1]$ defined by

$$f_\delta(x) = \mathbb{E}_{y \sim_\delta x} f(y) = \frac{1}{2\delta} \int_{x-\delta}^{x+\delta} f(y) \mathrm{d}y.$$

The function f_δ is the convolution of f and the uniform kernel $\phi : \mathbb{R} \to [0, 1]$ defined by $\phi(x) = \frac{1}{2\delta} \mathbf{1}[|x| \leq \delta]$.

Fix $\tau = \frac{4}{\epsilon} \mathbb{NS}_\delta(f)$. We introduce the function $g^* : [0, 1] \to \{0, 1, *\}$ by setting

$$g^*(x) = \begin{cases} 1 & \text{when } f_\delta(x) \geq 1 - \tau, \\ 0 & \text{when } f_\delta(x) \leq \tau, \text{ and} \\ * & \text{otherwise} \end{cases}$$

for all $x \in [0, 1]$. Finally, we define $g : [0, 1] \to \{0, 1\}$ by setting $g(x) = g^*(y)$ where $y \leq x$ is the largest value for which $g(y) \neq *$. (If no such y exists, we fix $g(x) = 0$.)

We first claim that $dist(f, g) \leq \frac{\epsilon}{2}$. To see this, note that

$$dist(f, g) = \Pr_x[f(x) \neq g(x)]$$

$$\leq \Pr_x[g^*(x) = *] + \Pr_x[f(x) = 0 \land g^*(x) = 1] + \Pr_x[f(x) = 1 \land g^*(x) = 0]$$

$$= \Pr_x[\tau < f_\delta(x) < 1 - \tau] + \Pr_x[f(x) = 0 \land f_\delta(x) \geq 1 - \tau] + \Pr_x[f(x) = 1 \land f_\delta(x) \leq \tau].$$

We bound the three terms on the RHS individually. For the first term, we observe that $\mathbb{NS}_\delta(f, x) = \min\{f_\delta(x), 1 - f_\delta(x)\}$ and that $\mathbb{E}_x \mathbb{NS}_\delta(f, x) = \mathbb{NS}_\delta(f)$. From these identities and Markov's inequality, we have that

$$\Pr_x[\tau < f_\delta(x) < 1 - \tau] = \Pr_x[\mathbb{NS}_\delta(f, x) > \tau] < \frac{\mathbb{NS}_\delta(f)}{\tau} = \frac{\epsilon}{4}.$$

For the second term, let $S \subseteq [0, 1]$ denote the set of points x where $f(x) = 0$ and $f_\delta(x) \geq 1 - \tau$. Let $\Gamma \subseteq S$ represent a δ-net of S. Clearly, $|\Gamma| \leq \frac{1}{\delta}$. For $x \in \Gamma$, let $B_x = (x - \delta, x + \delta)$ be a ball of radius δ around x. Since $f_\delta(x) \geq 1 - \tau$, the intersection of S and B_x has mass at most $|S \cap B_x| \leq \tau\delta$. Therefore, the total mass of S is at most $|S| \leq |\Gamma| \tau\delta = \tau$. By the bounds on the

33

noise sensitivity of f in the lemma's statement, we therefore have

$$\Pr_x[f(x) = 0 \wedge f_\delta(x) \geq 1 - \tau] \leq \tau \leq \tfrac{\epsilon}{8}.$$

Similarly, we obtain the same bound on the third term. As a result, $dist(f, g) \leq \tfrac{\epsilon}{4} + \tfrac{\epsilon}{8} + \tfrac{\epsilon}{8} = \tfrac{\epsilon}{2}$, as we wanted to show.

We now want to show that g is a union of $m \leq d\delta(1 + \tfrac{\epsilon}{2})$ intervals. Each left boundary of an interval in g occurs at a point $x \in [0, 1]$ where $g^*(x) = *$, where the maximum $y \leq x$ such that $g^*(y) \neq *$ takes the value $g^*(y) = 0$, and where the minimum $z \geq x$ such that $g^*(z) \neq *$ has the value $g^*(z) = 1$. In other words, for each left boundary of an interval in g, there exists an interval (y, z) such that $f_\delta(y) \leq \tau$, $f_\delta(z) \geq 1 - \tau$, and for each $y < x < z$, $f_\delta(x) \in (\tau, 1 - \tau)$. Fix any interval (y, z). Since f_δ is the convolution of f with a uniform kernel of width 2δ, it is Lipschitz continuous (with Lipschitz constant $\tfrac{1}{2\delta}$). So there exists $x \in (y, z)$ such that the conditions $f_\delta(x) = \tfrac{1}{2}$, $x - y \geq 2\delta(\tfrac{1}{2} - \tau)$, and $z - x \geq 2\delta(\tfrac{1}{2} - \tau)$ all hold. As a result,

$$\int_y^z \mathrm{NS}_\delta(f, t)\, \mathrm{d}t = \int_y^x \mathrm{NS}_\delta(f, t)\, \mathrm{d}t + \int_x^z \mathrm{NS}_\delta(f, t)\, \mathrm{d}t \geq 2\delta(\tfrac{1}{2} - \tau)^2.$$

Similarly, for each right boundary of an interval in g, we have an interval (y, z) such that

$$\int_y^z \mathrm{NS}_\delta(f, t)\, \mathrm{d}t \geq 2\delta(\tfrac{1}{2} - \tau)^2.$$

The intervals (y, z) for the left and right boundaries are all disjoints, so

$$\mathbb{NS}_\delta(f) \geq \sum_{i=1}^{2m} \int_{y^i}^{z^i} \mathrm{NS}_\delta(f, t)\, \mathrm{d}t \geq 2m \frac{\delta}{2}(1 - 2\tau)^2.$$

This means that

$$m \leq \frac{d\delta(1 + \epsilon/4)}{\delta(1 - 2\tau)^2} \leq d(1 + \tfrac{\epsilon}{2})$$

and g is a union of at most $d(1 + \tfrac{\epsilon}{2})$ intervals, as we wanted to show.

Finally, we want to show that any function that is the union of $m \leq d(1 + \tfrac{\epsilon}{2})$ intervals is $\tfrac{\epsilon}{2}$-close to a union of d intervals. Let $\ell_1, \ldots, \ell_m$ represent the lengths of the intervals in g. Clearly,

34

$\ell_1 + \cdots + \ell_m \leq 1$, so there must be a set S of $m - d \leq d\epsilon/2$ intervals in f with total length

$$\sum_{i \in S} \ell_i \leq \frac{m-d}{m} \leq \frac{d\epsilon/2}{d(1 + \frac{\epsilon}{2})} < \frac{\epsilon}{2}.$$

Consider the function $h : [0,1] \to \{0,1\}$ obtained by removing the intervals in S from g (i.e., by setting $h(x) = 0$ for the values $x \in [b_{2i-1}, b_{2i}]$ for some $i \in S$). The function h is a union of d intervals and $dist(g, h) \leq \frac{\epsilon}{2}$. This completes the proof, since $dist(f, h) \leq dist(f, g) + dist(g, h) \leq \epsilon$. $\qquad\square$

2.8 Proofs for Testing LTFs

We complete the proof that LTFs can be tested with $O(\sqrt{n})$ samples in this section.

For a fixed function $f : \mathbb{R}^n \to \mathbb{R}$, define $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ to be $g(x, y) = f(x)f(y) \langle x, y \rangle$. Let $g^* : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the truncation of g defined by setting

$$g^*(x, y) = \begin{cases} f(x)f(y) \langle x, y \rangle & \text{if } |\langle x, y \rangle| \leq \sqrt{4n \log(4n/\epsilon^3)} \\ 0 & \text{otherwise.} \end{cases}$$

Our goal is to estimate $\mathbb{E}g$. The following lemma shows that $\mathbb{E}g^*$ provides a good estimate of this value.

Lemma 2.22. *Let* $g, g^* : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *be defined as above. Then* $|\mathbb{E}g - \mathbb{E}g^*| \leq \frac{1}{2}\epsilon^3$.

Proof. For notational clarity, fix $\tau = \sqrt{4n \log(4n/\epsilon^3)}$. By the definition of g and g^* and with the trivial bound $|f(x)f(y) \langle x, y \rangle| \leq n$ we have

$$|\mathbb{E}g - \mathbb{E}g^*| = \left| \Pr_{x,y} \left[|\langle x, y \rangle| > \tau \right] \cdot \mathbb{E}_{x,y} \left[f(x)f(y) \langle x, y \rangle \mid |\langle x, y \rangle| > \tau \right] \right| \leq n \cdot \Pr_{x,y} \left[|\langle x, y \rangle| > \tau \right].$$

The right-most term can be bounded with a standard Chernoff argument. By Markov's inequality and the independence of the variables $x_1, \ldots, x_n, y_1, \ldots, y_n$,

$$\Pr_{x,y} \left[\langle x, y \rangle > \tau \right] = \Pr \left[e^{t\langle x,y \rangle} > e^{t\tau} \right] \leq \frac{\mathbb{E}e^{t\langle x,y \rangle}}{e^{t\tau}} = \frac{\prod_{i=1}^n \mathbb{E}e^{t x_i y_i}}{e^{t\tau}}.$$

The moment generating function of a standard normal random variable is $\mathbb{E}e^{ty} = e^{t^2/2}$, so

$$\mathbb{E}_{x_i, y_i}\left[e^{tx_i y_i}\right] = \mathbb{E}_{x_i}\left[\mathbb{E}_{y_i} e^{tx_i y_i}\right] = \mathbb{E}_{x_i} e^{(t^2/2)x_i^2}.$$

When $x \sim \mathcal{N}(0, 1)$, the random variable x^2 has a χ^2 distribution with 1 degree of freedom. The moment generating function of this variable is $\mathbb{E}e^{tx^2} = \sqrt{\frac{1}{1-2t}} = \sqrt{1 + \frac{2t}{1-2t}}$ for any $t < \frac{1}{2}$. Hence,

$$\mathbb{E}_{x_i} e^{(t^2/2)x_i^2} \leq \sqrt{1 + \frac{t^2}{1 - t^2}} \leq e^{\frac{t^2}{2(1-t^2)}}$$

for any $t < 1$. Combining the above results and setting $t = \frac{\tau}{2n}$ yields

$$\Pr_{x,y}\left[\langle x, y \rangle > \tau\right] \leq e^{\frac{nt^2}{2(1-t^2)} - t\tau} \leq e^{-\frac{\tau^2}{4n}} = \frac{\epsilon^3}{4n}.$$

The same argument shows that $\Pr[\langle x, y \rangle < -\tau] \leq \frac{\epsilon^3}{4n}$ as well. $\qquad\square$

The reason we consider the truncation g^* is that its smaller ℓ_∞ norm will enable us to apply a strong Bernstein-type inequality on the concentration of measure of the U-statistic estimate of $\mathbb{E}g^*$.

Lemma 2.23 (Arcones [Arcones, 1995]). *For a symmetric function $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, let $\Sigma^2 = \mathbb{E}_x[\mathbb{E}_y[h(x, y)]^2] - \mathbb{E}_{x,y}[h(x, y)]^2$, let $b = \|h - \mathbb{E}h\|_\infty$, and let $U_m(h)$ be a random variable obtained by drawing $x^1, \ldots, x^m$ independently at random and setting $U_m(h) = \binom{m}{2}^{-1} \sum_{i<j} h(x^i, x^j)$. Then for every $t > 0$,*

$$\Pr[|U_m(h) - \mathbb{E}h| > t] \leq 4\exp\left(\frac{mt^2}{8\Sigma^2 + 100bt}\right).$$

We are now ready to complete the proof of the upper bound of Theorem 2.5.

Theorem 2.24 (Upper bound in Theorem 2.5, restated). *Linear threshold functions can be tested over the standard n-dimensional Gaussian distribution with $O(\sqrt{n \log n})$ queries in both the active and passive testing models.*

Proof. Consider the LTF-TESTER algorithm. When the estimates $\tilde{\mu}$ and $\tilde{\nu}$ satisfy

$$|\tilde{\mu} - \mathbb{E}f| \leq \epsilon^3 \qquad \text{and} \qquad |\tilde{\nu} - \mathbb{E}[f(x)f(y)\langle x, y \rangle]| \leq \epsilon^3,$$

Lemmas 2.10 and 2.11 guarantee that the algorithm correctly distinguishes LTFs from functions that are far from LTFs. To complete the proof, we must therefore show that the estimates are within the specified error bounds with probability at least $2/3$.

The values $f(x^1), \ldots, f(x^m)$ are independent $\{-1, 1\}$-valued random variables. By Hoeffding's inequality,

$$\Pr[|\tilde{\mu} - \mathbb{E}f| \leq \epsilon^3] \geq 1 - 2e^{-\epsilon^6 m/2} = 1 - 2e^{-O(\sqrt{n})}.$$

The estimate $\tilde{\nu}$ is a U-statistic with kernel g^* as defined above. This kernel satisfies

$$\|g^* - \mathbb{E}g^*\|_\infty \leq 2\|g^*\|_\infty = 2\sqrt{4n \log(4n/\epsilon^3)}$$

and

$$\Sigma^2 \leq \mathbb{E}_y\big[\mathbb{E}_x[g^*(x, y)]^2\big] = \mathbb{E}_y\big[\mathbb{E}_x[f(x)f(y)\langle x, y\rangle \mathbf{1}[|\langle x, y\rangle| \leq \tau]]^2\big].$$

For any two functions $\phi, \psi : \mathbb{R}^n \to \mathbb{R}$, when ψ is $\{0, 1\}$-valued the Cauchy-Schwarz inequality implies that $\mathbb{E}_x[\phi(x)\psi(x)]^2 \leq \mathbb{E}_x[\phi(x)]\mathbb{E}_x[\phi(x)\psi(x)^2] = \mathbb{E}_x[\phi(x)]\mathbb{E}_x[\phi(x)\psi(x)]$ and so $\mathbb{E}_x[\phi(x)\psi(x)]^2 \leq \mathbb{E}_x[\phi(x)]$. Applying this inequality to the expression for Σ^2 gives

$$\Sigma^2 \leq \mathbb{E}_y\big[\mathbb{E}_x[f(x)f(y)\langle x, y\rangle]^2\big] = \mathbb{E}_y\Big[\Big(\sum_{i=1}^n f(y)y_i\mathbb{E}_x[f(x)x_i]\Big)^2\Big] = \sum_{i,j}\hat{f}(e_i)\hat{f}(e_j)\mathbb{E}_y[y_iy_j] = \sum_{i=1}^n \hat{f}(e_i)^2.$$

By Parseval's identity, we have $\sum_i \hat{f}(e_i)^2 \leq \|\hat{f}\|_2^2 = \|f\|_2^2 = 1$. Lemmas 2.22 and 2.23 imply that

$$\Pr[|\tilde{\nu} - \mathbb{E}g| \leq \epsilon^3] = \Pr[|\tilde{\nu} - \mathbb{E}g^*| \leq \tfrac{1}{2}\epsilon^3] \geq 1 - 4e^{-\frac{mt^2}{8 + 200\sqrt{n\log(4n/\epsilon^3)}t}} \geq \tfrac{11}{12}.$$

The union bound completes the proof of correctness. $\qquad\square$

2.9 Proofs for Testing Disjoint Unions

Theorem 2.12 (Restated). *Given properties $\mathcal{P}_1, \ldots, \mathcal{P}_N$, if each $\mathcal{P}_i$ is testable over D_i with $q(\epsilon)$ queries and $U(\epsilon)$ unlabeled samples, then their disjoint union $\mathcal{P}$ is testable over the combined distribution D with $O(q(\epsilon/2) \cdot (\log^3 \frac{1}{\epsilon}))$ queries and $O(U(\epsilon/2) \cdot (\frac{N}{\epsilon} \log^3 \frac{1}{\epsilon}))$ unlabeled samples.*

Proof. Let $p = (p_1, \ldots, p_N)$ denote the mixing weights for distribution D; that is, a random draw from D can be viewed as selecting i from distribution p and then selecting x from D_i. We are given that each $\mathcal{P}_i$ is testable with failure probability $1/3$ using using $q(\epsilon)$ queries and $U(\epsilon)$ unlabeled samples. By repetition, this implies that each is testable with failure probability δ using $q_\delta(\epsilon) = O(q(\epsilon) \log(1/\delta))$ queries and $U_\delta(\epsilon) = O(U(\epsilon) \log(1/\delta))$ unlabeled samples, where we will set $\delta = \epsilon^2$. We now test property $\mathcal{P}$ as follows:

For $\epsilon' = 1/2, 1/4, 1/8, \ldots, \epsilon/2$ do:

Repeat $O(\frac{\epsilon'}{\epsilon} \log(1/\epsilon))$ times:

1. Choose a random (i, x) from D.

2. Sample until either $U_\delta(\epsilon')$ samples have been drawn from D_i or $(8N/\epsilon)U_\delta(\epsilon')$ samples total have been drawn from D, whichever comes first.

3. In the former case, run the tester for property $\mathcal{P}_i$ with parameter ϵ', making $q_\delta(\epsilon')$ queries. If the tester rejects, then reject.

If all runs have accepted, then accept.

First to analyze the total number of queries and samples, since we can assume $q(\epsilon) \geq 1/\epsilon$ and $U(\epsilon) \geq 1/\epsilon$, we have $q_\delta(\epsilon')\epsilon'/\epsilon = O(q_\delta(\epsilon/2))$ and $U_\delta(\epsilon')\epsilon'/\epsilon = O(U_\delta(\epsilon/2))$ for $\epsilon' \geq \epsilon/2$. Thus, the total number of queries made is at most

$$\sum_{\epsilon'} q_\delta(\epsilon/2) \log(1/\epsilon) = O\left(q(\epsilon/2) \cdot \log^3 \frac{1}{\epsilon}\right)$$

and the total number of unlabeled samples is at most

$$\sum_{\epsilon'} \frac{8N}{\epsilon} U_\delta(\epsilon/2) \log(1/\epsilon) = O\left(U(\epsilon/2)\frac{N}{\epsilon} \log^3 \frac{1}{\epsilon}\right).$$

Next, to analyze correctness, if indeed $f \in \mathcal{P}$ then each call to a tester rejects with probability at most δ so the overall failure probability is at most $(\delta/\epsilon) \log^2(1/\epsilon) < 1/3$; thus it suffices to analyze the case that $dist_D(f, \mathcal{P}) \geq \epsilon$.

38

If $dist_D(f, \mathcal{P}) \geq \epsilon$ then $\sum_{i:p_i \geq \epsilon/(4N)} p_i \cdot dist_{D_i}(f_i, \mathcal{P}_i) \geq 3\epsilon/4$. Moreover, for indices i such that $p_i \geq \epsilon/(4N)$, with high probability Step 2 draws $U_\delta(\epsilon')$ samples, so we may assume for such indices the tester for $\mathcal{P}_i$ is indeed run in Step 3. Let $I = \{i : p_i \geq \epsilon/(4N) \text{ and } dist_{D_i}(f_i, \mathcal{P}_i) \geq \epsilon/2\}$. Thus, we have

$$\sum_{i \in I} p_i \cdot dist_{D_i}(f_i, \mathcal{P}_i) \geq \epsilon/4.$$

Let $I_{\epsilon'} = \{i \in I : dist_{D_i}(f_i, \mathcal{P}_i) \in [\epsilon', 2\epsilon']\}$. Bucketing the above summation by values ϵ' in this way implies that for some value $\epsilon' \in \{\epsilon/2, \epsilon, 2\epsilon, \dots, 1/2\}$, we have:

$$\sum_{i \in I_{\epsilon'}} p_i \geq \epsilon/(8\epsilon' \log(1/\epsilon)).$$

This in turn implies that with probability at least $2/3$, the run of the algorithm for this value of ϵ' will find such an i and reject, as desired. $\qquad\square$

2.10 Proofs for Testing Dimensions

2.10.1 Passive Testing Dimension (proof of Theorem 2.15)

Lower bound: By design, $d_{passive}$ is a lower bound on the number of examples needed for passive testing. In particular, if $D_S(\pi, \pi') \leq 1/4$, and if the target is with probability $1/2$ chosen from π and with probability $1/2$ chosen from π', even the Bayes optimal tester will fail to identify the correct distribution with probability $\frac{1}{2} \sum_{y \in \{0,1\}^{|S|}} \min(\pi_S(y), \pi'_S(y)) = \frac{1}{2}(1 - D_S(\pi, \pi')) \geq 3/8$. The definition of $d_{passive}$ implies that there exist $\pi \in \Pi_0$, $\pi' \in \Pi_\epsilon$ such that $\Pr_S(D_S(\pi, \pi') \leq 1/4) \geq 3/4$. Since π' has a $1 - o(1)$ probability mass on functions that are ϵ-far from $\mathcal{P}$, this implies that over random draws of S and f, the overall failure probability of any tester is at least $(1 - o(1))(3/8)(3/4) > 1/4$. Thus, at least $d_{passive} + 1$ random labeled examples are required if we wish to guarantee error at most $1/4$. This in turn implies $\Omega(d_{passive})$ examples are needed to guarantee error at most $1/3$.

Upper bound: We now argue that $O(d_{passive})$ examples are *sufficient* for testing as well. Toward this end, consider the following natural testing game. The adversary chooses a function f such that either $f \in \mathcal{P}$ or $dist_D(f, \mathcal{P}) \geq \epsilon$. The tester picks a function A that maps labeled samples of size k to accept/reject. That is, A is a deterministic passive testing algorithm. The payoff to the tester is the probability that A is correct when S is chosen iid from D and labeled by f.

If $k > d_{passive}$ then (by definition of $d_{passive}$) we know that for any distribution π over $f \in \mathcal{P}$ and any distribution π' over f that are ϵ-far from $\mathcal{P}$, we have $\Pr_{S \sim D^k}(\mathrm{D}_S(\pi, \pi') > 1/4) > 1/4$. We now need to translate this into a statement about the value of the game. The key fact we can use is that if the adversary uses distribution $\alpha\pi + (1 - \alpha)\pi'$ (i.e., with probability α it chooses from π and with probability $1 - \alpha$ it chooses from π'), then the Bayes optimal predictor has error exactly

$$\sum_y \min(\alpha\pi_S(y), (1 - \alpha)\pi'_S(y)) \leq \max(\alpha, 1 - \alpha) \sum_y \min(\pi_S(y), \pi'_S(y)),$$

while

$$\sum_y \min(\pi_S(y), \pi'_S(y)) = 1 - (1/2) \sum_y |\pi_S(y) - \pi'_S(y)| = 1 - \mathrm{D}_S(\pi, \pi'),$$

so that the Bayes risk is at most $\max(\alpha, 1 - \alpha)(1 - \mathrm{D}_S(\pi, \pi'))$. Thus, for any $\alpha \in [7/16, 9/16]$, if $\mathrm{D}_S(\pi, \pi') > 1/4$, the Bayes risk is less than $(9/16)(3/4) = 27/64$. Furthermore, any $\alpha \notin [7/16, 9/16]$ has Bayes risk at most $7/16$. Thus, since $\mathrm{D}_S(\pi, \pi') > 1/4$ with probability $> 1/4$ (and if $\mathrm{D}_S(\pi, \pi') \leq 1/4$ then the error probability of the Bayes optimal predictor is at most $1/2$), for any mixed strategy of the adversary, the Bayes optimal predictor has risk less than $(1/4)(7/16) + (3/4)(1/2) = 31/64$.

Now, applying the minimax theorem we get that for $k = d_{passive} + 1$, there exists a mixed strategy A for the tester such that for any function chosen by the adversary, the probability the tester is correct is at least $1/2 + \gamma$ for a constant $\gamma > 0$ (namely, $1/64$). We can now boost the correctness probability using a constant-factor larger sample. Specifically, let $m = c \cdot (d_{passive} + 1)$ for some constant c, and consider a sample S of size m. The tester simply partitions the sample

S into c pieces, runs A separatately on each piece, and then takes majority vote. This gives us that $O(d_{passive})$ examples are sufficient for testing with any desired constant success probability in $(1/2, 1)$.

2.10.2 Coarse Active Testing Dimension (proof of Theorem 2.17)

Lower bound: First, we claim that any nonadaptive active testing algorithm that uses $\leq d_{coarse}/c$ label requests must use more than n^c unlabeled examples (and thus no algorithm can succeed using $o(d_{coarse})$ labels). To see this, suppose algorithm A draws n^c unlabeled examples. The number of subsets of size d_{coarse}/c is at most $n^{d_{coarse}}/6$ (for $d_{coarse}/c \geq 3$). So, by definition of d_{coarse} and the union bound, with probability at least $5/6$, *all* such subsets S satisfy the property that $D_S(\pi, \pi') < 1/4$. Therefore, for any sequence of such label requests, the labels observed will not be sufficient to reliably distinguish π from π'. Adaptive active testers can potentially choose their next point to query based on labels observed so far, but the above immediately implies that even adaptive active testers cannot use an $o(\log(d_{coarse}))$ queries.

Upper bound: For the upper bound, we modify the argument from the passive testing dimension analysis as follows. We are given that for any distribution π over $f \in \mathcal{P}$ and any distribution π' over f that are ϵ-far from $\mathcal{P}$, for $k = d_{coarse}+1$, we have $\Pr_{S \sim D^k}(D_S(\pi, \pi') > 1/4) > n^{-k}$. Thus, we can sample $U \sim D^m$ with $m = \Theta(k \cdot n^k)$, and partition U into subsamples $S_1, S_2, \ldots, S_{cn^k}$ of size k each. With high probability, at least one of these subsamples S_i will have $D_S(\pi, \pi') > 1/4$. We can thus simply examine each subsample, identify one such that $D_S(\pi, \pi') > 1/4$, and query the points in that sample. As in the proof for the passive bound, this implies that for any strategy for the adversary in the associated testing game, the best response has probability at least $1/2+\gamma$ of success for some constant $\gamma > 0$. By the minimax theorem, this implies a testing strategy with success probability $1/2+\gamma$ which can then be boosted to $2/3$. The total number of label requests used in the process is only $O(d_{coarse})$.

Note, however, that this strategy uses a number of unlabeled examples $\Omega(n^{d_{coarse}+1})$. Thus,

41

this only implies an active tester for $d_{coarse} = O(1)$. Nonetheless, combining the upper and lower bounds yields Theorem 2.17.

2.10.3 Active Testing Dimension (proof of Theorem 2.19)

Lower bound: for a given sample U, we can think of an adaptive active tester as a decision tree, defined based on which example it would request the label of next given that the previous requests have been answered in any given way. A tester making k queries would yield a decision tree of depth k. By definition of $d_{active}(u)$, with probability at least $3/4$ (over choice of U), any such tester has error probability at least $(1/4)(1 - o(1))$ over the choice of f. Thus, the overall failure probability is at least $(3/4)(1/4)(1 - o(1) > 1/8$.

Upper bound: We again consider the natural testing game. We are given that for any mixed strategy of the adversary with equal probability mass on functions in $\mathcal{P}$ and functions ϵ-far from $\mathcal{P}$, the best response of the tester has expected payoff at least $(1/4)(3/4) + (3/4)(1/2) = 9/16$. This in turn implies that for any mixed strategy at all, the best response of the tester has expected payoff at least $33/64$ (if the adversary puts more than $17/32$ probability mass on either type of function, the tester can just guess that type with expected payoff at least $17/32$, else it gets payoff at least $(1 - 1/16)(9/16) > 33/64$). By the minimax theorem, this implies existence of a randomized strategy for the tester with at least this payoff. We then boost correctness using $c \cdot u$ samples and $c \cdot d_{active}(u)$ queries, running the tester c times on disjoint samples and taking majority vote.

2.10.4 Lower Bounds for Testing LTFs (proof of Theorem 2.20)

We complete the proofs for the lower bounds on the query complexity for testing linear threshold functions in the active and passive models. This proof has three parts. First, in Section 2.10.4, we introduce some preliminary (technical) results that will be used to prove the lower bounds on the passive and coarse dimensions of testing LTFs. In Section 2.10.4, we introduce some more pre-

liminary results regarding random matrices that we will use to bound the active dimension of the class. Finally, in Section 2.10.4, we put it all together and complete the proof of Theorem 2.20.

Preliminaries for $d_{passive}$ and d_{coarse}

Fix any K. Let the dataset $X = \{x_1, x_2, \cdots, x_K\}$ be sampled iid according to the uniform distribution on $\{-1, +1\}^n$ and let $\mathbf{X} \in \mathcal{R}^{K \times n}$ be the corresponding data matrix.

Suppose $\mathbf{w} \sim N(0, I_{n \times n})$. We let

$$\mathbf{z} = \mathbf{X}\mathbf{w},$$

and note that the conditional distribution of $\mathbf{z}$ given X is normal with mean 0 and (X-dependent) covariance matrix, which we denote by Σ. Further applying threshold function to $\mathbf{z}$ gives $\mathbf{y}$ as the predicted label vector of an LTF.

Lemma 2.25. *For any matrix B, $\log(det(B)) = Tr(\log(B))$, where $\log(B)$ is the matrix exponential of B.*

Proof. From [Higham, 2008], we know since every eigenvalue of A corresponds to the eigenvalue of $\exp(A)$, thus

$$det(\exp(A)) = \exp(Tr(A)) \tag{2.1}$$

where $\exp(A)$ is the matrix exponential of A. Taking logarithm of both sides of (2.1), we get

$$\log(det(\exp(A))) = Tr(A) \tag{2.2}$$

Let $B = \exp(A)$ (thus $A = \log(B)$). Then (2.2) can rewritten as $\log(det(B)) = Tr(\log B)$. $\qquad\square$

Lemma 2.26. *For sufficiently large n, and a value $K = \Omega(\sqrt{n/\log(K/\delta)})$, with probability at least $1 - \delta$ (over X),*

$$\|\mathbb{P}_{(\mathbf{z}/\sqrt{n})|X} - N(0, I)\| \leq 1/4.$$

Proof. Let l be the feature index. For a pair $\mathbf{x}_i$ and $\mathbf{x}_j$,

$$\mathbb{P}(\left|\left|\{l : x_{il} = x_{jl}\}\right| - \frac{n}{2}\right| > \sqrt{\frac{n \log \frac{2}{\delta}}{2}}) \leq \delta$$

By Hoeffding Inequality, with probability $1 - \delta$,

$$
\begin{aligned}
\mathbf{x}_i^T \mathbf{x}_j &= |\{l : x_{il} = x_{jl}\}| - |\{l : x_{il} \neq x_{jl}\}| \\
&= 2|\{l : x_{il} = x_{jl}\}| - n \in \left[-2\sqrt{\frac{n \log \frac{2}{\delta}}{2}}, 2\sqrt{\frac{n \log \frac{2}{\delta}}{2}}\right]
\end{aligned}
$$

By union bound,

$$\mathbb{P}\left(\exists i, j, \text{ such that } \mathbf{x}_i^T \mathbf{x}_j \notin \left[-\sqrt{2n \log \frac{2K^2}{\delta}}, \sqrt{2n \log \frac{2K^2}{\delta}}\right]\right) \leq K^2 \frac{\delta}{K^2} = \delta \qquad (2.3)$$

For the remainder of the proof we suppose the (probability $1 - \delta$) event $\forall i, j, \mathbf{x}_i^T \mathbf{x}_j \in \left[-\sqrt{2n \log(2K^2/\delta)}, \sqrt{2n \log(2K^2/\delta)}\right]$ occurs.

$$
\begin{aligned}
Cov(z_i/\sqrt{n}, z_j/\sqrt{n}|X) &= \frac{\mathbb{E}[z_i z_j | X]}{n} \\
&= \frac{1}{n}\mathbb{E}\left[(\sum_{l=1}^{n} w_l \cdot x_{il})(\sum_{l=1}^{n} w_l \cdot x_{jl})|X\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{l,m=1,1}^{n,n} w_l w_m x_{il} x_{jm}|X\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_l w_l^2 x_{il} x_{jl}|X\right] = \frac{1}{n}\mathbb{E}\left[\sum_l x_{il} x_{jl}|X\right] \\
&= \frac{1}{n}\sum_l x_{il} x_{jl} = \frac{1}{n}\mathbf{x}_i^T \mathbf{x}_j \in \left[-\sqrt{\frac{2 \log(2K^2/\delta)}{n}}, \sqrt{\frac{2 \log(2K^2/\delta)}{n}}\right]
\end{aligned}
$$

because $\mathbb{E}[w_l w_m] = 0$ (for $l \neq m$) and $\mathbb{E}[w_l^2] = 1$. Let $\beta = \sqrt{\frac{2 \log(2K^2/\delta)}{n}}$. Thus Σ is a $K \times K$ matrix, with $\Sigma_{ii} = 1$ for $i = 1, \cdots, K$ and $\Sigma_{ij} \in [-\beta, \beta]$ for all $i \neq j$.

Let $P_1 = N(0, \Sigma^{K \times K})$ and $P_2 = N(0, I^{K \times K})$. As the density

$$p_1(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^K \det(\Sigma)}} \exp(-\frac{1}{2}\mathbf{z}^T \Sigma^{-1} \mathbf{z})$$

44

and the density

$$p_2(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^K}} \exp(-\frac{1}{2}\mathbf{z}^T\mathbf{z})$$

Then L_1 distance between the two distributions P_1 and P_2

$$|dP_2 - dP_1| \leq 2\sqrt{K(P_1, P_2)} = 2\sqrt{(1/2)\log\det(\Sigma)},$$

where this last equality is by [Davis and Dhillon, 2006]. By Lemma 2.25, $\log(\det(\Sigma)) = Tr(\log(\Sigma))$. Write $A = \Sigma - I$. By the Taylor series

$$\log(I + A) = -\sum_{i=1}^{\infty} \frac{1}{i}(I - (I+A))^i = -\sum_{i=1}^{\infty} \frac{1}{i}(-A)^i$$

$$\text{Thus} \quad Tr(\log(I + A)) = \sum_{i=1}^{\infty} \frac{1}{i}Tr((-A)^i). \tag{2.4}$$

Every entry in A^i can be expressed as a sum of at most K^{i-1} terms, each of which can be expressed as a product of exactly i entries from A. Thus, every entry in A^i is in the range $[-K^{i-1}\beta^i, K^{i-1}\beta^i]$. This means $Tr(A^i) \leq K^i\beta^i$. Therefore, if $K\beta < 1/2$, since $Tr(A) = 0$, the expansion of $Tr(\log(I + A)) \leq \sum_{i=2}^{\infty} K^i\beta^i = O\left(K^2\frac{\log(K/\delta)}{n}\right)$.

In particular, for some $K = \Omega(\sqrt{n/\log(K/\delta)})$, $Tr(\log(I+A))$ is bounded by the appropriate constant to obtain the stated result. $\qquad\square$

Preliminaries for d_{active}

Given an $n \times m$ matrix A with real entries $\{a_{i,j}\}_{i\in[n],j\in[m]}$, the *adjoint* (or *transpose* – the two are equivalent since A contains only real values) of A is the $m \times n$ matrix A^* whose (i,j)-th entry equals $a_{j,i}$. Let us write $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ to denote the eigenvalues of $\sqrt{A^*A}$. These values are the *singular values* of A. The matrix A^*A is positive semidefinite, so the singular values of A are all non-negative. We write $\lambda_{\max}(A) = \lambda_1$ and $\lambda_{\min}(A) = \lambda_m$ to represent its largest and smallest singular values. Finally, the *induced norm* (or *operator norm*) of A is

$$\|A\| = \max_{x\in\mathbb{R}^m\setminus\{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x\in\mathbb{R}^m:\|x\|_2^2=1} \|Ax\|_2.$$

45

For more details on these definitions, see any standard linear algebra text (e.g., [Shilov, 1977]). We will also use the following strong concentration bounds on the singular values of random matrices.

Lemma 2.27 (See [Vershynin, 2012, Cor. 5.35]). *Let A be an $n \times m$ matrix whose entries are independent standard normal random variables. Then for any $t > 0$, the singular values of A satisfy*

$$\sqrt{n} - \sqrt{m} - t \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq \sqrt{n} + \sqrt{m} + t \tag{2.5}$$

with probability at least $1 - 2e^{-t^2/2}$.

The proof of this lemma follows from Talagrand's inequality and Gordon's Theorem for Gaussian matrices. See [Vershynin, 2012] for the details. The lemma implies the following corollary which we will use in the proof of our theorem.

Corollary 2.28. *Let A be an $n \times m$ matrix whose entries are independent standard normal random variables. For any $0 < t < \sqrt{n} - \sqrt{m}$, the $m \times m$ matrix $\frac{1}{n} A^* A$ satisfies both inequalities*

$$\left\| \tfrac{1}{n} A^* A - I \right\| \leq 3 \frac{\sqrt{m} + t}{\sqrt{n}} \quad and \quad \det\left(\tfrac{1}{n} A^* A \right) \geq e^{-m\left(\frac{(\sqrt{m}+t)^2}{n} + 2 \frac{\sqrt{m}+t}{\sqrt{n}} \right)} \tag{2.6}$$

with probability at least $1 - 2e^{-t^2/2}$.

Proof. When there exists $0 < z < 1$ such that $1 - z \leq \frac{1}{\sqrt{n}} \lambda_{\max}(A) \leq 1 + z$, the identity $\frac{1}{\sqrt{n}} \lambda_{\max}(A) = \| \frac{1}{\sqrt{n}} A \| = \max_{\|x\|_2^2 = 1} \| \frac{1}{\sqrt{n}} A x \|_2$ implies that

$$1 - 2z \leq (1 - z)^2 \leq \max_{\|x\|_2^2 = 1} \left\| \tfrac{1}{\sqrt{n}} A x \right\|_2^2 \leq (1 + z)^2 \leq 1 + 3z.$$

These inequalities and the identity $\| \frac{1}{n} A^* A - I \| = \max_{\|x\|_2^2 = 1} \| \frac{1}{\sqrt{n}} A x \|_2^2 - 1$ imply that $-2z \leq \| \frac{1}{n} A^* A - I \| \leq 3z$. Fixing $z = \frac{\sqrt{m}+t}{\sqrt{n}}$ and applying Lemma 2.27 completes the proof of the first inequality.

Recall that $\lambda_1 \leq \cdots \leq \lambda_m$ are the eigenvalues of $\sqrt{A^* A}$. Then

$$\det(\tfrac{1}{n} A^* A) = \frac{\det(\sqrt{A^* A})^2}{n} = \frac{(\lambda_1 \cdots \lambda_m)^2}{n} \geq \left(\frac{\lambda_1^2}{n} \right)^m = \left(\frac{\lambda_{\min}(A)^2}{n} \right)^m.$$

Lemma 2.27 and the elementary inequality $1 + x \leq e^x$ complete the proof of the second inequality. □

Proof of Theorem 2.20

Theorem 2.20 (Restated). *For linear threshold functions under the uniform distribution on* $\{-1, 1\}^n$, $d_{passive} = \Omega(\sqrt{n/\log(n)})$ *and* $d_{active} = \Omega((n/\log(n))^{1/3})$.

Proof. Let K be as in Lemma 2.26 for $\delta = 1/4$. Let $D = \{(x_1, y_1), \ldots, (x_K, y_K)\}$ denote the sequence of labeled data points under the random LTF based on $\mathbf{w}$. Furthermore, let $D' = \{(x_1, y_1'), \ldots, (x_K, y_K')\}$ denote the sequence of labeled data points under a target function that assigns an independent random label to each data point. Also let $\mathbf{z}_i = (1/\sqrt{n})\mathbf{w}^T x_i$, and let $\mathbf{z}' \sim N(0, I_{K \times K})$. Let $E = \{(x_1, \mathbf{z}_1), \ldots, (x_K, \mathbf{z}_K)\}$ and $E' = \{(x_1, \mathbf{z}_1'), \ldots, (x_K, \mathbf{z}_K')\}$. Note that we can think of y_i and y_i' as being functions of $\mathbf{z}_i$ and $\mathbf{z}_i'$, respectively. Thus, letting $X = \{x_1, \ldots, x_K\}$, by Lemma 2.26, with probability at least $3/4$,

$$\|\mathbb{P}_{D|X} - \mathbb{P}_{D'|X}\| \leq \|\mathbb{P}_{E|X} - \mathbb{P}_{E'|X}\| \leq 1/4.$$

This suffices for the claim that $d_{passive} = \Omega(K) = \Omega(\sqrt{n/\log(n)})$.

Next we turn to the lower bound on d_{active}. Let us now introduce two distributions $\mathcal{D}_{\text{yes}}$ and $\mathcal{D}_{\text{no}}$ over linear threshold functions and functions that (with high probability) are far from linear threshold functions, respectively. We draw a function f from $\mathcal{D}_{\text{yes}}$ by first drawing a vector $\mathbf{w} \sim \mathcal{N}(0, I_{n \times n})$ from the n-dimensional standard normal distribution. We then define $f : x \mapsto \text{sgn}(\frac{1}{\sqrt{n}} x \cdot \mathbf{w})$. To draw a function g from $\mathcal{D}_{\text{no}}$, we define $g(x) = \text{sgn}(\mathbf{y}_x)$ where each $\mathbf{y}_x$ variable is drawn independently from the standard normal distribution $\mathcal{N}(0, 1)$.

Let $\mathbf{X} \in \mathbb{R}^{n \times q}$ be a random matrix obtained by drawing q vectors from the n-dimensional normal distribution $\mathcal{N}(0, I_{n \times n})$ and setting these vectors to be the columns of $\mathbf{X}$. Equivalently, $\mathbf{X}$ is the random matrix whose entries are independent standard normal variables. When we view $\mathbf{X}$ as a set of q queries to a function $f \sim \mathcal{D}_{\text{yes}}$ or a function $g \sim \mathcal{D}_{\text{no}}$, we get $f(\mathbf{X}) = \text{sgn}(\frac{1}{\sqrt{n}}\mathbf{X}\mathbf{w})$

and $g(\mathbf{X}) = \mathrm{sgn}(\mathbf{y_X})$. Note that $\frac{1}{\sqrt{n}}\mathbf{Xw} \sim \mathcal{N}(0, \frac{1}{n}\mathbf{X}^*\mathbf{X})$ and $\mathbf{y_X} \sim \mathcal{N}(0, I_{q \times q})$. To apply Lemma 2.21 it suffices to show that the ratio of the pdfs for both these random variables is bounded by $\frac{6}{5}$ for all but $\frac{1}{5}$ of the probability mass.

The pdf $p : \mathbb{R}^q \to \mathbb{R}$ of a q-dimensional random vector from the distribution $\mathcal{N}_{q \times q}(0, \Sigma)$ is

$$p(x) = (2\pi)^{-\frac{q}{2}} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}x^T\Sigma^{-1}x}.$$

Therefore, the ratio function $r : \mathbb{R}^q \to \mathbb{R}$ between the pdfs of $\frac{1}{\sqrt{n}}\mathbf{Xw}$ and of $\mathbf{y_X}$ is

$$r(x) = \det(\tfrac{1}{n}\mathbf{X}^*\mathbf{X})^{-\frac{1}{2}} e^{\frac{1}{2}x^T((\frac{1}{n}\mathbf{X}^*\mathbf{X})^{-1}-I)x}.$$

Note that

$$x^T((\tfrac{1}{n}\mathbf{X}^*\mathbf{X})^{-1} - I)x \le \|(\tfrac{1}{n}\mathbf{X}^*\mathbf{X})^{-1} - I\|\|x\|_2^2 = \|\tfrac{1}{n}\mathbf{X}^*\mathbf{X} - I\|\|x\|_2^2,$$

so by Lemma 2.27 with probability at least $1 - 2e^{-t^2/2}$ we have

$$r(x) \le e^{\frac{q}{2}\left(\frac{(\sqrt{q}+t)^2}{n} + 2\frac{\sqrt{q}+t}{\sqrt{n}}\right) + 3\frac{\sqrt{q}+t}{\sqrt{n}}\|x\|_2^2}.$$

By a union bound, for $U \sim \mathcal{N}(0, I_{n \times n})^u$, $u \in \mathbb{N}$ with $u \ge q$, the above inequality for $r(x)$ is true for all subsets of U of size q, with probability at least $1 - u^q 2e^{-t^2/2}$. Fix $q = n^{\frac{1}{3}}/(50(\ln(u))^{\frac{1}{3}})$ and $t = 2\sqrt{q\ln(u)}$. Then $u^q 2e^{-t^2/2} \le 2u^{-q}$, which is $< 1/4$ for any sufficiently large n. When $\|x\|_2^2 \le 3q$ then for large n, $r(x) \le e^{74/625} < \frac{6}{5}$. To complete the proof, it suffices to show that when $x \sim \mathcal{N}(0, I_{q \times q})$, the probability that $\|x\|_2^2 > 3q$ is at most $\frac{1}{5}2^{-q}$. The random variable $\|x\|_2^2$ has a χ^2 distribution with q degrees of freedom and expected value $\mathbb{E}\|x\|_2^2 = \sum_{i=1}^q \mathbb{E}x_i^2 = q$. Standard concentration bounds for χ^2 variables imply that

$$\Pr_{x \sim \mathcal{N}(0, I_{q \times q})}[\|x\|_2^2 > 3q] \le e^{-\frac{4}{3}q} < \tfrac{1}{5}2^{-q},$$

as we wanted to show. Thus, Lemma 2.21 implies $\mathrm{err}^*(\mathrm{DT}_q, \mathrm{Fair}(\pi, \pi', U)) > 1/4$ holds whenever this $r(x)$ inequality is satisfied for all subsets of U of size q; we have shown this happens with probabiliity greater than $3/4$, so we must have $d_{active} \ge q$. $\qquad\square$

If we are only interested in bounding d_{coarse}, the proof can be somewhat simplified. Specifically, taking $\delta = n^{-K}$ in Lemma 2.26 implies that with probability at least $1 - n^{-K}$,

$$\|\mathbb{P}_{D|X} - \mathbb{P}_{D'|X}\| \leq \|\mathbb{P}_{E|X} - \mathbb{P}_{E'|X}\| \leq 1/4,$$

which suffices for the claim that $d_{coarse} = \Omega(K)$, where $K = \Omega(\sqrt{n/K \log(n)})$: in particular, $d_{coarse} = \Omega((n/\log(n))^{1/3})$.

2.11 Testing Semi-Supervised Learning Assumptions

We now consider testing of common assumptions made in semi-supervised learning [Chapelle, Schlkopf, and Zien, 2006], where unlabeled data, together with assumptions about how the target function and data distribution relate, are used to constrain the search space. As mentioned in Section 2.4, one such assumption we can test using our generic disjoint-unions tester is the cluster assumption, that if data lies in N identifiable clusters, then points in the same cluster should have the same label. We can in fact achieve the following tighter bounds:

Theorem 2.29. *We can test the cluster assumption with active testing using $O(N/\epsilon)$ unlabeled examples and $O(1/\epsilon)$ queries.*

Proof. Let p_{i1} and p_{i0} denote the probability mass on positive examples and negative examples respectively in cluster i, so $p_{i1} + p_{i0}$ is the total probabilty mass of cluster i. Then $dist(f, \mathcal{P}) = \sum_i \min(p_{i1}, p_{i0})$. Thus, a simple tester is to draw a random example x, draw a random example y from x's cluster, and check if $f(x) = f(y)$. Notice that with probability *exactly $dist(f, \mathcal{P})$*, point x is in the minority class of its own cluster, and conditioned on this event, with probability at least $1/2$, point y will have a different label. It thus suffices to repeat this process $O(1/\epsilon)$ times. One complication is that as stated, this process might require a large *unlabeled* sample, especially if x belongs to a cluster i such that $p_{i0} + p_{i1}$ is small, so that many draws are needed to find a point y in x's cluster. To achieve the given *unlabeled* sample bound, we initially draw an unlabeled sample of size $O(N/\epsilon)$ and simply perform the above test on the uniform distribution

49

U over that sample, with distance parameter $\epsilon/2$. Standard sample complexity bounds [Vapnik, 1998] imply that $O(N/\epsilon)$ unlabeled points are sufficient so that if $dist_D(f, \mathcal{P}) \geq \epsilon$ then with high probability, $dist_U(f, \mathcal{P}) \geq \epsilon/2$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

We now consider the property of a function having a large margin with respect to the underlying distribution: that is, the distribution D and target f are such that any point in the support of $D|_{f=1}$ is at distance γ or more from any point in the support of $D|_{f=0}$. This is a common property assumed in graph-based and nearest-neighbor-style semi-supervised learning algorithms [Chapelle, Schlkopf, and Zien, 2006]. Note that we are not additionally requiring the target to be a linear separator or have any special functional form. For scaling, we assume that points lie in the unit ball in R^d, where we view d as constant and $1/\gamma$ as our asymptotic parameter.[8] Since we are not assuming any specific functional form for the target, the number of labeled examples needed for *learning* could be as large as $\Omega(1/\gamma^d)$ by having a distribution with support over $\Omega(1/\gamma^d)$ points that are all at distance γ from each other (and therefore can be labeled arbitrarily). Furthermore, passive testing would require $\Omega(1/\gamma^{d/2})$ samples as this specific case encodes the cluster-assumption setting with $N = \Omega(1/\gamma^d)$ clusters. We will be able to perform active testing using only $O(1/\epsilon)$ label requests.

First, one distinction between this and other properties we have been discussing is that it is a property of the *relation* between the target function f and the distribution D; i.e., of the combined distribution $D_f = (D, f)$ over labeled examples. As a result, the natural notion of *distance* to this property is in terms of the variation distance of D_f to the closest D_* satisfying the property.[9] Second, we will have to also allow some amount of slack on the γ parameter as

---

[8]Alternatively points could lie in a d-dimensional manifold in some higher-dimensional ambient space, where the property is defined with respect to the manifold, and we have sufficient unlabeled data to "unroll" the manifold using existing methods [Chapelle, Schlkopf, and Zien, 2006, Roweis and Saul, 2000, Tenenbaum, Silva, and Langford, 2000].

[9]As a simple example illustrating the issue, consider $X = [0, 1]$, a target f that is negative on $[0, 1/2)$ and positive on $[1/2, 1]$, and a distribution D that is uniform but where the region $[1/2, 1/2 + \gamma]$ is downweighted to

well. Specifically, our tester will distinguish the case that D_f indeed has margin γ from the case that the D_f is ϵ-far from having margin γ' where $\gamma' = \gamma(1 - 1/c)$ for some constant $c > 1$; e.g., think of $\gamma' = \gamma/2$. This slack can also be seen to be necessary (see discussion following the proof of Theorem 2.13). In particular, we have the following.

Theorem 2.13 (Restated). *For any γ, $\gamma' = \gamma(1 - 1/c)$ for constant $c > 1$, for data in the unit ball in R^d for constant d, we can distinguish the case that D_f has margin γ from the case that D_f is ϵ-far from margin γ' using Active Testing with $O(1/(\gamma^{2d}\epsilon^2))$ unlabeled examples and $O(1/\epsilon)$ label requests.*

Proof. First, partition the input space X (the unit ball in R^d) into regions $R_1, R_2, \ldots, R_N$ of diameter at most $\gamma/(2c)$. By a standard volume argument, this can be done using $N = O(1/\gamma^d)$ regions (absorbing "c" into the $O()$). Next, we run the cluster-property tester on these N regions, with distance parameter $\epsilon/4$. Clearly, if the cluster-tester rejects, then we can reject as well. Thus, we may assume below that the total impurity within individual regions is at most $\epsilon/4$.

Now, consider the following weighted graph G_γ. We have N vertices, one for each of the N regions. We have an edge (i, j) between regions R_i and R_j if $\mathrm{diam}(R_i \cup R_j) < \gamma$. We define the *weight* $w(i, j)$ of this edge to be $\min(D[R_i], D[R_j])$ where $D[R]$ is the probability mass in R under distribution D. Notice that if there is no edge between region R_i and R_j, then by the triangle inequality every point in R_i must be at distance at least γ' from every point in R_j. Also, note that each vertex has degree $O(c^d) = O(1)$, so the total weight over all edges is $O(1)$. Finally, note that while algorithmically we do not know the edge weights precisely, we can estimate all edge weights to $\pm\epsilon/(4M)$, where $M = O(N)$ is the total number of edges, using the unlabeled sample size bounds given in the Theorem statement. Let $\tilde{w}(i, j)$ denote the estimated weight of edge (i, j).

Let $E_{witness}$ be the set of edges (i, j) such that one endpoint is majority positive and one is

have total probability mass only $1/2^n$. Such a D_f is $1/2^n$-close to the property under variation distance, but would be nearly $1/2$-far from the property if the only operation allowed were to change the function f.

majority negative. Note that if D_f satisfies the γ-margin property, then every edge in $E_{witness}$ has weight 0. On the other hand, if D_f is ϵ-far from the γ'-margin property, then the total weight of edges in $E_{witness}$ is at least $3\epsilon/4$. The reason is that otherwise one could convert D_f to D'_f satisfying the margin condition by zeroing out the probability mass in the lightest endpoint of every edge $(i,j) \in E_{witness}$, and then for each vertex, zeroing out the probability mass of points in the minority label of that vertex. (Then, renormalize to have total probability 1.) The first step moves distance at most $3\epsilon/4$ and the second step moves distance at most $\epsilon/4$ by our assumption of success of the cluster-tester. Finally, if the true total weight of edges in $E_{witness}$ is at least $3\epsilon/4$ then the sum of their estimated weights $\tilde{w}(i,j)$ is at least $\epsilon/2$. This implies we can perform our test as follows. For $O(1/\epsilon)$ steps, do:

1. Choose an edge (i,j) with probability proportional to $\tilde{w}(i,j)$.

2. Request the label for a random $x \in R_i$ and $y \in R_j$. If the two labels disagree, then reject.

If D_f is ϵ-far from the γ'-margin property, then each step has probability $\tilde{w}(E_{witness})/\tilde{w}(E) = O(\epsilon)$ of choosing a witness edge, and conditioned on choosing a witness edge has probability at least $1/2$ of detecting a violation. Thus, overall, we can test using $O(1/\epsilon)$ labeled examples and $O(1/(\gamma^{2d}\epsilon^2))$ unlabeled examples. $\qquad\square$

On the necessity of slack in testing the margin assumption: Consider an instance space $\mathcal{X} = [0,1]^2$ and two distributions over labeled examples D_1 and D_2. Distribution D_1 has probability mass $1/2^{n+1}$ on positive examples at location $(0, i/2^n)$ and negative examples at $(\gamma', i/2^n)$ for each $i = 1, 2, \ldots, 2^n$, for $\gamma' = \gamma(1 - 1/2^{2n})$. Notice that D_1 is $1/2$-far from the γ-margin property because there is a matching between points in the support of $D_1|_{f=1}$ and points in the support of $D_1|_{f=0}$ where the matched points have distance less than γ. On the other hand, for each $i = 1, 2, \ldots, 2^n$, distribution D_2 has probability mass $1/2^n$ at either a positive point $(0, i/2^n)$ or a negative point $(\gamma', i/2^n)$, chosen at random, but zero probability mass at the other location. Distribution D_2 satisfies the γ-margin property, and yet D_1 and D_2 cannot be distinguished using

a polynomial number of unlabeled examples.

Chapter 3

Testing Piecewise Real-Valued Functions

Abstract

This chapter extends the model of the previous chapter to the setting of testing properties of real-valued functions. Specifically, it establishes a technique for testing d-piecewise constantness of a real-valued function.

3.1   Piecewise Constant

For this section, let $\mathbb{NS}_\delta = \mathbb{NS}_\delta^1 = \int_0^1 \mathbb{NS}_\delta^1(x)\mathrm{d}x$, where $\mathbb{NS}_\delta^1(x) = \frac{1}{2\delta}\int_{x-\delta}^{x+\delta} \mathbb{I}[f(x) \neq f(y)]\mathrm{d}y$.

Proposition 3.1. *Fix $\delta > 0$ and let $f : [0,1] \to \mathbb{R}$ be a d-piecewise constant function. Then* $\mathbb{NS}_\delta(f) \leq (d-1)\frac{\delta}{2}$.

Proof. For any fixed $b \in [0,1]$, the probability that $x < b < y$ when $x \sim U(0,1)$ and $y \sim U(x-\delta, x+\delta)$ is

$$\Pr_{x,y}[x < b < y] = \int_0^\delta \Pr_{y \sim U(b-t-\delta, b-t+\delta)}[y \geq b]\mathrm{d}t = \int_0^\delta \frac{\delta - t}{2\delta}\mathrm{d}t = \frac{\delta}{4}.$$

Similarly, $\Pr_{x,y}[y < b < x] = \frac{\delta}{4}$. So the probability that b lies between x and y is at most $\frac{\delta}{2}$.

When f is a d-piecewise constant function, $f(x) \neq f(y)$ only if at least one of the boundaries $b_1, \ldots, b_{d-1}$ of the regions of f lie in between x and y. So by the union bound, $\Pr[f(x) \neq$

$f(y)] \leq (d-1)(\delta/2)$. Note that if b is within distance δ of 0 or 1, the probability is only lower. $\qquad\square$

Lemma 3.2. *Fix* $\delta = \frac{\epsilon^2}{32d}$. *Let* $f : [0,1] \to \mathbb{R}$ *be any function with noise sensitivity* $\mathbb{NS}_\delta(f) \leq (d-1)\frac{\delta}{2}(1+\frac{\epsilon}{4})$. *Then* f *is* ϵ-*close to a* d-*piecewise constant function.*

Proof. The proof proceeds in two steps: We first show that f is $\frac{\epsilon}{2}$-close to a $(1+(d-1)(1+\frac{\epsilon}{2}))$-piecewise constant function, and then we show that every $(1+(d-1)(1+\frac{\epsilon}{2}))$-piecewise constant function is $\frac{\epsilon}{2}$-close to a d-piecewise constant function.

For eacy $y \in \mathbb{R}$, consider the function $f_\delta^y : [0,1] \to [0,1]$ defined by

$$f_\delta^y(x) = \frac{1}{2\delta} \int_{x-\delta}^{x+\delta} \mathbb{I}[f(t) = y]\mathrm{d}t.$$

The function f_δ^y is the convolution of $f^y = \mathbb{I}[f = y]$ and the uniform kernel $\phi : \mathbb{R} \to [0,1]$ defined by $\phi(x) = \frac{1}{2\delta}\mathbf{1}[|x| \leq \delta]$.

Note that for any x, there is at most one value $y \in \mathbb{R}$ for which $f_\delta^y(x) > 1/2$. Fix $\tau = \frac{4}{\epsilon}\mathbb{NS}_\delta(f)$. We introduce the function $g^* : [0,1] \to \mathbb{R} \cup \{*\}$ by setting

$$g^*(x) = \begin{cases} \mathrm{argmax}_{y\in\mathbb{R}} f_\delta^y(x) & \text{when } \sup_{y\in\mathbb{R}} f_\delta^y(x) \geq 1 - \tau, \\ * & \text{otherwise} \end{cases}$$

for all $x \in [0,1]$. Finally, we define $g : [0,1] \to \{0,1\}$ by setting $g(x) = g^*(z)$ where $z \leq x$ is the largest value for which $g^*(z) \neq *$. (If no such z exists, we let $g(x) = g^*(z)$ for the smallest value $z \geq x$ with $g^*(z) \neq *$; if that does not exist, then for completeness define $g(x) = 0$ everywhere, though this case will not come up).

We first claim that $dist(f,g) \leq \frac{\epsilon}{4}$. To see this, note that

$$dist(f,g) = \Pr_x[f(x) \neq g(x)]$$

$$\leq \Pr_x[g^*(x) = *] + \Pr_x[* \neq g^*(x) \neq f(x)]$$

$$= \Pr_x[\sup_{y\in\mathbb{R}} f_\delta^y(x) < 1 - \tau] + \Pr_x[\sup_{y\in\mathbb{R}\backslash\{f(x)\}} f_\delta^y(x) \geq 1 - \tau].$$

Because $\tau < 1/2$, at most one y can have $f_\delta^y(x) \geq 1 - \tau$, so that both $\sup_{y \in \mathbb{R}} f_\delta^y(x) < 1 - \tau$ and $\sup_{y \in \mathbb{R} \setminus \{f(x)\}} f_\delta^y(x) \geq 1 - \tau$ imply $f_\delta^{f(x)}(s) < 1 - \tau$; thus, since these events are disjoint, the above sum of probabilities is at most

$$\Pr_x[f_\delta^{f(x)}(x) < 1 - \tau].$$

Now observe that $\mathrm{NS}_\delta(f, x) = 1 - f_\delta^{f(x)}(x)$ and that $\mathbb{E}_x \mathrm{NS}_\delta(f, x) = \mathbb{NS}_\delta(f)$. From these identities and Markov's inequality, we have that

$$\Pr_x[f_\delta^{f(x)}(x) < 1 - \tau] = \Pr_x[1 - f_\delta^{f(x)}(x) > \tau] = \Pr_x[\mathrm{NS}_\delta(f, x) > \tau] < \frac{\mathbb{NS}_\delta(f)}{\tau} = \frac{\epsilon}{4}.$$

We now want to show that g is m-piecewise constant, for some $m \leq d(1 + \frac{\epsilon}{2})$. Since each f_δ^y is the convolution of $\mathbb{I}[f = y]$ with a uniform kernel of width 2δ, it is Lipschitz continuous (with Lipschitz constant $\frac{1}{2\delta}$). Also recall that $\tau < 1/2$, and at most one value y can have $f_\delta^y(x) \geq 1 - \tau$ for any given x. Thus, if we consider any two points $x, z \in [0, 1]$ with $* \neq g^*(x) \neq g^*(z) \neq *$ and $x < z$, it must be that $|x - z| \geq 2\delta 2(\frac{1}{2} - \tau)$, and that there is at least one point $t \in (x, z)$ with $\sup_{y \in \mathbb{R}} f_\delta^y(t) = 1/2$. Since each f_δ^y is $\frac{1}{2\delta}$-Lipschitz, so is $\sup_{y \in \mathbb{R}} f_\delta^y$, so that we have

$$\int_{t - 2\delta(\frac{1}{2} - \tau)}^{t + 2\delta(\frac{1}{2} - \tau)} f_\delta^{f(s)}(s) \mathrm{d}s \leq \int_{t - 2\delta(\frac{1}{2} - \tau)}^{t + 2\delta(\frac{1}{2} - \tau)} \sup_{y \in \mathbb{R}} f_\delta^y(s) \mathrm{d}s$$
$$\leq 2 \int_0^{2\delta(\frac{1}{2} - \tau)} (\frac{1}{2} + \frac{s}{2\delta}) \mathrm{d}s = 2\delta(\frac{1}{2} - \tau)(\frac{3}{2} - \tau).$$

Therefore,

$$\int_x^z \mathrm{NS}_\delta(f, s) \mathrm{d}s = \int_x^z (1 - f_\delta^{f(s)}(s)) \mathrm{d}s \geq (z - x) - 2\delta(\frac{1}{2} - \tau)(\frac{3}{2} - \tau)$$
$$\geq 2\delta 2(\frac{1}{2} - \tau) - 2\delta(\frac{1}{2} - \tau)(\frac{3}{2} - \tau) = 2\delta(\frac{1}{2} - \tau)(\frac{1}{2} + \tau)) = 2\delta(\frac{1}{4} - \tau^2).$$

Since any x with $g^*(x) \neq *$ has $g(x) = g^*(x)$, and since g is defined to be continuous from the right on $[0, 1]$, for every transition point $x > 0$ for g (i.e., a point x for which there exist arbitrarily close points z having $g(z) \neq g(x)$), there is a point $z < x$ such that every $t \in (z, x)$

56

has $g(t) = g^*(z) \neq g^*(x) = g(x)$; combined with the above, we have that $\int_z^x \text{NS}_\delta(f, s) \mathrm{d}s \geq 2\delta(\frac{1}{4} - \tau^2)$. Altogether, if g has m such transition points, then

$$\mathbb{NS}_\delta(f) = \int_0^1 \text{NS}_\delta(f, s)\mathrm{d}s \geq m2\delta(\frac{1}{4} - \tau^2).$$

By assumption, $\mathbb{NS}_\delta(f) \leq (d-1)\frac{\delta}{2}(1 + \frac{\epsilon}{4})$. Therefore, we must have

$$m \leq \frac{(d-1)\delta(1 + \frac{\epsilon}{4})}{4\delta(\frac{1}{4} - \tau^2)} \leq (d-1)\frac{1 + \frac{\epsilon}{4}}{1 - 4\tau^2} \leq (d-1)\frac{1 + \frac{\epsilon}{4}}{(1 - 2\tau)^2} \leq (d-1)(1 + \frac{\epsilon}{2}).$$

In particular, this means g is $(m+1)$-piecewise constant, for an $m \leq (d-1)(1 + \frac{\epsilon}{2})$.

Finally, we want to show that any $(m+1)$-piecewise constant function, for $m \leq (d-1)(1+\frac{\epsilon}{2})$, is $\frac{\epsilon}{2}$-close to a d-piecewise constant function. Let $\ell_1, \ldots, \ell_{m+1}$ represent the lengths of the m regions in g. Clearly, $\ell_1 + \cdots + \ell_{m+1} = 1$, so there must be a set S of $(m+1) - d \leq (d-1)\epsilon/2$ regions in g with total length

$$\sum_{i \in S} \ell_i \leq \frac{(m+1) - d}{(m+1)} \leq \frac{(d-1)\epsilon/2}{1 + (d-1)(1 + \frac{\epsilon}{2})} < \frac{\epsilon}{2}.$$

Consider the function $h : [0, 1] \to \{0, 1\}$ obtained by removing the regions in S from g (i.e., for each x in a region indexed by $i \in S$, setting $h(x) = h(z)$ for z a point in the nearest region to x that is not indexed by some $j \in S$). The function h is then d-piecewise constant, and $dist(g, h) \leq \frac{\epsilon}{2}$. This completes the proof, since $dist(f, h) \leq dist(f, g) + dist(g, h) \leq \epsilon$. $\square$

With these results, applying the same technique as used in the unions of intervals method in the previous chapter yields a tester for d-piecewise constant functions.

Chapter 4

Learnability of DNF with Representation-Specific Queries

Abstract

[1]We study the problem of PAC learning the space of DNF functions with a type of query specific to the representation of the target DNF. Specifically, given a pair of positive examples from a polynomial-sized sample, our query asks whether the two examples satisfy a term in common in the target DNF. We show that a number of interesting special types of DNF targets are efficiently properly learnable with this type of query, though the general problem of learning an arbitrary DNF target under an arbitrary distribution is no easier than in the traditional PAC model. Specifically, we find that 2-term DNF are efficiently properly learnable under arbitrary distributions, as are disjoint DNF. We further study the special case of learning under the uniform distribution, and find that several other general families of DNF functions are efficiently properly learnable with these queries, including functions with $O(\log(n))$ relevant variables, and monotone DNF functions for which each variable appears in at most $O(\log(n))$ terms.

We also study a variety of generalizations of this type of query. For instance, consider in-

---

[1]Joint work with Avrim Blum and Jaime Carbonell.

stead the ability to ask how many terms a pair of examples satisfy in common, where the examples are again taken from a polynomial-sized sample. In this case, we can efficiently properly learn several more general classes of DNF, including DNF having $O(\log(n))$ terms, DNF having $O(\log(n))$ relevant variables, DNF for which each example can satisfy at most $O(1)$ terms, all under arbitrary distributions. Other possible generalizations of the query include allowing the algorithm to ask the query for an arbitrary number of examples from the sample at once (rather than just two), or allowing the algorithm to ask the query for examples of its own construction; we show that both of these generalizations allow for efficient proper learnability of arbitrary DNF functions under arbitrary distributions.

4.1 Introduction

Consider a bank aiming to use machine learning to identify instances of financial fraud. To do so, the bank would have experts label past transactions as fraudulent or not, and then run a learning algorithm on the resulting labeled data. However, this learning problem might be quite difficult because of the existence of multiple intrinsic types of fraud, with each positive example perhaps involving multiple types. That is, the target might be a DNF formula, a class for which no efficient algorithms are known.

Yet in such a case, perhaps the experts performing the labeling could be called on to provide a bit more information. In particular, suppose that given two positive examples of fraud, the experts could indicate whether or not the two examples are *similar* in the sense of having at least one intrinsic type of fraud (at least one term) in common. Or perhaps the experts could indicate *how* similar the examples are (how many terms in common they satisfy). This is certainly substantially more information. Can it be used to learn DNF formulas and their natural subclasses efficiently?

In our work, we study the problem of learning DNF formulas and other function classes using such pairwise, representation-dependent queries. Specifically, we consider queries of the form, "Do these two positive examples satisfy at least one term in common in the target DNF

formula?" (we call these *boolean similarity queries*) and "How many terms in common do these two positive examples satisfy?" (we call these *numerical similarity queries*).

4.1.1 Our Results

We begin with a somewhat surprising negative result, that learning general DNF formulas under arbitrary distributions from boolean similarity queries is as hard as PAC-learning DNF formulas without them. This result uses the equivalence between group learning, weak learning, and strong learning. In contrast, learning disjoint DNF (a class that contains decision trees) with such queries is quite easy. We in addition show that it helps in a number of other important cases, including properly learning "parsimonious" DNF formulas (formulas for which no term can be deleted without appreciably changing the function) as well as any 2-term DNF, a class known to be NP-Hard to properly learn from labeled data alone.

Under the uniform distribution, we can properly learn any DNF formula for which each variable appears in $O(\log(n))$ terms, as well as any DNF formula with $O(\log(n))$ relevant variables.

If we are allowed to ask numerical similarity queries, then we show we can properly learn any DNF formula having $O(\log(n))$ terms, under arbitrary distributions, or any DNF formula having $O(\log(n))$ relevant variables, again under arbitrary distributions. If we are allowed to ask "Do these k examples satisfy any term in common?" for arbitrary (poly-sized) k, we can even properly learn arbitrary DNF formulas under arbitrary distributions.

This topic of learning with representation-specific queries is interesting, even beyond the DNF case, and we have explored a variety of other learning problems of this type as well.

4.2 Learning DNF with General Queries: Hardness Results

Theorem 4.1. *Learning DNF from random data under arbitrary distributions with boolean similarity queries is as hard as learning DNF from random data under arbitrary distributions with*

only the labels (no queries).

Proof. [Kearns, 1989] and [Kearns, Li, and Valiant, 1994] proved that "group learning" is equivalent to "weak learning".

In group learning, at each round we are given $poly(n)$ examples that are either all iid from $D+$ or all iid from $D-$ (i.e. all positive or all negative) and our goal is to figure out which case it is. Later, of course, Schapire [Schapire, 1990] proved that weak-learning is equivalent to strong-learning. So, if DNF is hard to PAC-learn, then DNF is also hard to group-learn.

Now, consider the following reduction from group-learning DNF in the standard model to learning DNF in the extended queries model. In particular, given an algorithm $\mathring{A}$ for learning from a polynomial number of examples in the extended queries model, we show how to use $\mathring{A}$ to group-learn as follows:

Given a set S of $m = poly(n)$ examples x_1, x_2, ..., x_m (we will use $m = tn$ where t is the number of terms in the target), construct a new example by just concatenating them together. So overall we now have nm variables. We present this concatenated example to $\mathring{A}$ with label equal to the label of S. If $\mathring{A}$ makes a similarity query between two positive examples $[x_1, x_2, ..., x_m]$ and $[x'_1, x'_2, ..., x'_m]$, we simply output *yes* (i.e., that they do indeed share a term in common).

We now argue that with high probability, the labels and our responses to $\mathring{A}$ are all fully consistent with some DNF formula of size mt. In particular, we claim they will be consistent with a target function that is just the AND of m copies of the original target function.

First of all, note that the AND of m copies of the original target function will produce the correct labels since by assumption either all $x_i \in S$ are positive or all $x_i \in S$ are negative. Next, we claim that whp, any two of these concatenated positive examples will share a term in common. Specifically, if the original DNF formula has t terms, then for two random positive examples from $D+$ there is probability at least $1/t$ that they share a common term. So, the chance of failure for two concatenated examples is at most $(1 - 1/t)^m$. (Because the only way that two of these big concatenated examples $[x_1, x_2, ..., x_m]$ and $[x'_1, x'_2, ..., x'_m]$ can fail to share a term in

common is if x_1 and x_1' fail, x_2 and x_2' fail, etc.). Setting $m = tn$, the probability of failure for any given query is at most $1/e^n$. Applying the union bound over all polynomially-many pairs of positive examples in Å's sample yields that with high probability all our responses are consistent. Therefore, by assumption, Å will produce a low-error hypothesis under the distribution over concatenated examples, which yields a low-error hypothesis for the group-learning problem. $\square$

We can extend the above result to "approximate numerical" queries that give the correct answer up to $1 \pm \tau$ for some constant $\tau > 0$ (or even $\tau \geq 1/poly(n)$).

Theorem 4.2. *Learning DNF from random data under arbitrary distributions with approximate-numerical-valued queries is as hard as learning DNF from random data under arbitrary distributions with only the labels (no queries).*

Proof. Assume we have an algorithm $\mathcal{A}$ that learns to error $\epsilon/2$ given a similarity oracle that tells us how many terms two examples have in common, up to a multiplicative factor τ. Specifically, if C is the number of terms in common, the oracle returns a value in the range $[(1-\tau)C, (1+\tau)C]$.

Now we do the reduction from group learning as before, forming higher-dimensional examples by concatenating groups $x_1, \cdots, x_m$, all of the same class, but this time with $m = 2n(t^4)(1 + \tau/2)^2/\tau^2$. Suppose, for now, that we know for the original DNF formula, the expected number of terms α that two that two random positive examples would have in common (we discharge this assumption later). In that case, when queried by Å for the similarity between two positive examples x, x', we simply answer with the closest integer to αm. As before, we argue that with high probability, our answers are consistent with a DNF formula g consisting of just m shifted copies of the original DNF.

Note that for a random pair of the concatenatedl examples composed of positive sub-examples, the expected number of terms in common in g is $m\alpha$. Furthermore, the number of terms in common is a sum of m independent samples of the original random variable (the one with mean α), each of which is bounded in the range $[0, t]$. So Hoeffding's inequality implies that with probability $1 - 2e^{-2m^2\alpha^2(\tau/2)^2/(m(t^2)(1+\tau/2)^2)} = 1 - 2e^{-n}$ (since $\alpha \geq 1/t$), the number C of terms in com-

mon satisfies $|C - m\alpha| \leq m\alpha(\tau/2)/(1+\tau/2)$, which implies $(1-\tau/2)C \leq m\alpha \leq (1+\tau/2)C$.

Thus, for a $poly(n)$-sized sample of data points, with high probability, all of the pairs of positive concatenated examples have the nearest integer to $m\alpha$ within these factors of their true number of terms in common. It therefore suffices to respond to $\mathcal{A}$'s similarity queries with the nearest integer to $m\alpha$.

Now the only trouble is that we do not know α. So we just try all positive integers i from 1 to mt and then use a validation set to select among the hypotheses produced. That is, we run $\mathcal{A}$ on the constructed data set and respond to all similarity queries with a single value i, getting back a classifier for these concatenated examples, and then repeat for each i. Then we take $O((1/\epsilon)\log(mt/\delta))$ additional higher-dimensional samples (with labels) and choose the classifier among these mt returned classifiers, having the smallest number of mistakes there-on. At least one of these mt values of i is the closest integer to $m\alpha$, so at least one of these mt classifiers is $\epsilon/2$-good, and our validation set will identify one whose error is at most ϵ. So we can use this classifier to identify whether a random m-sized group of examples is composed of all positives or all negatives, with error rate epsilon: i.e., we can do group learning.

If the algorithm A only has a "high probability" guarantee on success, we can repeat this several times with independent data sets, to boost the confidence that there will be a good classifier among those we choose from at the end, and slightly increase the size of the validation set to compensate for this larger number of classifiers. $\qquad\square$

4.3   Learning DNF with General Queries : Positive

4.3.1   Methods

The Neighborhood Method

We refer to the following simple procedure as the "neighborhood method". Take $m = \text{poly}(n, 1/\epsilon, \log(1/\delta))$ samples. First, among the positive examples, query all pairs (with the binary-valued query) to

construct a graph, in which examples are adjacent if they satisfy a term in common. For each positive example, construct a minimal conjunction consistent with that example and all of its neighbors (i.e., the consistent conjunction having largest number of literals in it). Next, discard any of these conjunctions that make mistakes on any negative examples. Then sequentially remove any conjunction c_1 such that some other remaining conjunction c_2 subsumes it (contains a subset of the variables). Form a DNF from the remaining conjunctions. Produce this resultant DNF as the output hypothesis.

Lemma 4.3. *Suppose the target DNF has $t = \text{poly}(n)$ terms. For an appropriate (t-dependent) polynomial sample size m, the neighborhood method will, with probability at least $1 - delta$, produce an ϵ-accurate DNF if, for each term T_i in the target DNF having a probability of satisfaction at least $\epsilon/2t$, there is at least a $p = 1/\text{poly}(n, 1/\epsilon)$ probability that a random example satisfies term T_i and no other term (we call such an example a "nice seed" for T_i).*

Proof. Under these conditions, $m = O((1/p)\log(t/\delta) + (t/\epsilon)\log(1/\epsilon\delta))$ samples suffice to guarantee each T_i with probability of satisfaction at least $\epsilon/2t$ has at least one nice seed, with probability at least $1 - \delta/2$.

In the second phase, we remove any conjunction inconsistent with the negative examples. The conjunctions guarnateed by the above argument survive this pruning due to their minimality, and the fact that they are learned from a set of examples that actually are consistent with some term in the target DNF (due to the nice seed). The final pruning step, which removes any redundancies in the set of conjunctions, leaves at most t conjunctions.

The terms that do not have nice seeds compose at most $\epsilon/2$ total probability mass, and m is large enough so that with probability at least $1 - \delta/4$, at most an $\epsilon/4$-fraction of the data satisfy these terms. Thus, since the result of the neighborhood method is a DNF formula with at most t terms, which correctly labels a $1 - \epsilon/2$ fraction of the m examples, the standard PAC bounds imply that with probability at least $1 - \delta/4$, the resulting DNF has error rate at most ϵ. A union bound over the above events implies this holds with probability at least $1 - \delta$. $\qquad\square$

The Common Profile Approach

In the case of numerical queries, we have some additional flexibility in designing a method. In this context, we refer to the following procedure as the "common profiles approach".

Consider a sample of $m = \text{poly}(n, 1/\epsilon, \log(1/\delta))$ random labeled examples, and for each pair of positive examples x, y, we request the number $K(x, y)$ of terms they satisfy in common; we additionally request $K(x, x)$ for each positive example x. For each positive example x, we identify the set S of examples y such that the numerical value of $K(x, y)$ is equal $K(x, x)$. So these points satisfy at least all the terms x satisfies. For each such set S, we learn a minimal conjunction consistent with these examples. Then for each of these conjunctions, if it is a specialization of some other one of the conjunctions, we discard it. Then we form our hypothesis DNF with the remaining conjunctions as the terms.

For any example x, relative to a particular target DNF, we refer to the "profile" of x as the set of terms T_i in the target DNF satisfied by x.

Lemma 4.4. *If the target DNF has at most $p = \text{poly}(n)$ possibel profiles, then the common profile approach, with an appropriate (p-dependent) sample size m, will with probability at least $1 - \delta$, produce a DNF having error rate at most ϵ.*

Proof. Note that this procedure produces a DNF that correctly labels the entire data set, since $K(x, y) = K(x, x)$ implies x and y have the same profiles, so that in particular the set S has some term in common to all the examples. If there are only a $\text{poly}(n)$ number of possible profiles, then the above will only produce at most as many distinct terms in its hypothesis DNF, so that a sufficiently large $\text{poly}(n)$-sized data set will be sufficient to guarantee good generalization error. Specifically, $m = O((pn/\epsilon) \log(1/\epsilon\delta))$ examples are enough to guarantee with probability at least $1 - \delta$, any DNF consistent with the data having at most p terms will have error rate at most ϵ, so this is sufficient for the common profile approach. $\qquad\square$

4.3.2 Positive Results

Theorem 4.5. *With numerical-valued queries, we can properly learn any DNF having $O(\log(n))$ relevant variables, under arbitrary distributions.*

Proof. These targets have $\mathrm{poly}(n)$ possible profiles, so the common profiles approach will be successful. $\square$

Theorem 4.6. *If the target DNF has only $O(\log(n))$ terms, then we can efficiently properly learn from random data under any distribution using numerical-valued queries.*

Proof. There are only $\mathrm{poly}(n)$ number of possible profiles, so the "common profiles" approach will work. $\square$

The above result is interesting particularly because proper learning (even for 2-term DNF) is known to be hard from labeled data alone.

Theorem 4.7. *If the target DNF has $t = \mathrm{poly}(n)$ terms, and is such that any example can satisfy at most $O(1)$ terms, then we can efficiently properly learn from random data using numerical-valued queries.*

Proof. There are at most $\mathrm{poly}(t) = \mathrm{poly}(n)$ possible profiles, so the "common profiles" approach will work. $\square$

Corollary 4.8. *We can properly learn any k-term DNF with numerical-valued queries, where k is constant.*

Proof. This follows from either Theorem 4.6 or Theorem 4.7. $\square$

Corollary 4.9. *If the DNF is such that any example can satisfy at most 1 term (a so-called "disjoint" DNF), then we can efficiently properly learn from random data using binary-valued queries.*

Proof. A numerical query whose value can be at most 1 is just a binary query anyway. $\square$

In particular, Decision Trees can be thought of as a DNF where each example satisfies at most 1 term.

Lemma 4.10. *If it happens that the target DNF is parsimonious (no redundant terms) for some random $\Omega((tn/\epsilon)\log(1/\epsilon) + (1/\epsilon)\log(1/\delta))$-sized data set (for any distribution), then we can efficiently produce a DNF consistent with it having at most t terms using binary-valued queries.*

Proof. (Sketch) Parsimonious, in this case, means that we cannot remove any terms without changing some labels. But this means that every term has some example that satisfies only that term (i.e., a nice seed). So as described in the proof of Lemma 4.3 above, the "neighborhood method," produces a DNF with terms for the neighborhoods of each of these nice seeds, which in the parsimonious case, covers all of the positive examples. □

Theorem 4.11. *We can properly learn 2-term DNF with binary queries.*

Proof. Take $O((n/\epsilon)\log(1/\epsilon) + (1/\epsilon)\log(1/\delta))$ random labeled examples and make the binary query for all pairs of positive examples. First, find a minimal conjunction consistent with all of the positive examples; if this conjunction does not misclassify any negative examples, return it. By classic PAC bounds, a conjunction consistent with this many random labeled examples will, with probabiliy at least $1 - \delta$, have error rate at most ϵ. Otherwise, if this conjunction misclassifies some negatives, then we are assured the target DNF is parsimonious for this data set, and thus Lemma 4.10 guarantees we can efficiently identify a 2-term DNF consistent with it using the binary-valued queries. Again, the classic PAC bounds imply the sample size is large enough to, with probability at least $1 - \delta$, guarantee that any consistent 2-term DNF has error rate at most ϵ. □

Theorem 4.11 gives a concrete result where using this type of query overturns a known hardness result for supervised learning.

Open problem Can this idea be extended to learning 3-term DNF or higher, still using only the binary-valued queries? Or is there a hardness result for properly learning 3-term DNF with

these binary-valued pairwise queries?

4.4   Learning DNF under the Uniform Distribution

In this section, we investigate the problem of learning DNF under a uniform distribution on $\{0,1\}^n$, using the binary-valued queries.

Definition 4.12. *Fix a constant $c \in (0,\infty)$. We say a term t in the target DNF is "relatively distinct" if it contains a variable v which occurs in at most $c\log(n)$ other terms. We say v is a witness to t being relatively distinct.*

Definition 4.13. *For a term t in the target DNF, and a variable v in t, we say v is "sometimes nonredundant" for t if, given a random example that satisfies t, there is at least an ϵ probability that every term in the target DNF that the example satisfies also contains v.*

Theorem 4.14. *Suppose no term in the target DNF is logically entailed from any other term in the target DNF, every term t is relatively distinct, and that some variable v that is a witness to t being relatively distinct is sometimes nonredundant for t. Then we can properly learn any monotone DNF of this type under a uniform distribution on $\{0,1\}^n$ with binary pairwise queries.*

Proof. By Lemma 4.3, it suffices to show that every term having at least $\epsilon/(2T)$ probability of being satisfied will, with high probability, have some example satisfying only that term, given a polynomial-sized data set.

Consider a given term t in the target DNF, and choose the v that witnesses relative distinctness which is sometimes nonredundant. Note that every other term in the target DNF contains some variable not present in t, and in particular this is true for the (at most) $c\log(n)$ terms containing v. So under the conditional distribution given that t is satisfied and that v is nonredundant, with probability at least $2^{-c\log(n)} = n^{-c}$, none of these other terms containing v are satisfied, so that t is the only term satisfied. Thus, since t has probability at least $\epsilon/(2T)$ of being satisfied, and v has probability at least ϵ of being nonredundant given that t is satisfied, we have that with probability

68

at least $(\epsilon^2/T)n^{-c}$, a random example satisfies t and no other terms in the target DNF.

Since this is the case for all terms in the target, a sample of size $O((T/\epsilon^2)n^c \log(T/\delta))$ guarantees every term has some example satisfying only that term, with probability at least $1 - \delta$. $\square$

We can also consider the class of DNF function having only a small number of relevant variables. In this context, it is interesting to observe that if the ith variable is irrelevant, then $P(K(x, y) = 1$ and $x_i \neq y_i) = P(K(x, y) = 1$ and $x_i = y_i)$, where x and y are independent uniformly-distributed samples, and $K(x, y) = 1$ iff x and y are positive examples that satisfy at least one term in common. However, as the following lemma shows, this is not true for relevant variables.

Lemma 4.15. *For x and y independent uniformly-distributed samples, if the target function has r relevant variables, and the ith variable is relevant in the target function, then $P(K(x, y) = 1$ and $x_i = y_i) - P(K(x, y) = 1$ and $x_i \neq y_i) \geq (1/4)^r$.*

Proof. For each pair (x, y) with $x_i \neq y_i$, there is a unique corresponding pair (x', y) with $x'_j = x_j$ for $j \neq i$, and $x'_i = y_i$. Let M_i be the number of x, y pairs with $x_i \neq y_i$ and $K(x, y) = 1$. Then note that for every x, y pair with $x_i \neq y_i$ and $K(x, y) = 1$, we also have $K(x', y) = 1$, since whatever term x and y satisfy in common cannot contain variable i anyway, so flipping that feature in x does not change whether x and y share a term or not. In particular, this implies the number of x, y pairs with $x_i = y_i$ and $K(x, y) = 1$ is at least M_i. However, we can also argue it is strictly larger, as follows. By definition of "relevant", each of the 2^r settings of the relevant variables corresponds to an equivalence class of feature vectors, all of which have the same label, and if that label is positive, then all of which have the same profile. Since variable i is relevant, at least one of the 2^r settings of the relevant variables yields an equivalence class of positive examples whose profile contains only terms with variable i in them (these are positive examples such that flipping variable i makes them negative). The probability that both x and y (chosen at random) are in this equivalence class is $(1/4)^r$. Note that for the (x, y) pairs of this type, we have $K(x, y) = 1$; however, if we flip feature x_i, then x would become negative, and

hence $K(x, y)$ would no longer be 1; this means this (x, y) pair is not included among those M_i pairs constructed above by flipping x_i starting from some (x, y) with $x_i \neq y_i$ and $K(x, y) = 1$. So $P(K(x, y) = 1$ and $x_i = y_i) - P(K(x, y) = 1$ and $x_i \neq y_i) = (M_i/4^n + (1/4)^r) - M_i/4^n = (1/4)^r$. $\qquad\square$

Theorem 4.16. *Under the uniform distribution, with binary pairwise queries, we can properly learn any DNF having $O(\log(n))$ relevant variables.*

Proof. We can use the property in Lemma 4.15 to design an algorithm as follows. For each i, sample $\Omega(8^r \log(n/\delta))$ random pairs (x, y), and evaluate $K(x, y)$ for each pair. Then calculate the difference of empirical probabilities (fraction of pairs (x, y) for which $K(x, y) = 1$ and $x_i = y_i$ minus fraction of pairs (x, y) for which $K(x, y) = 1$ and $x_i \neq y_i$). If this difference is $> (1/2)(1/4)^r$, decide variable i is relevant, and otherwise decide variable i is irrelevant. By Hoeffding and union bounds, with probability $1 - \delta/2$, this will find exactly the r relevant variables. Now enumerate all $2^r = \text{poly}(n)$ possible conjunctions that can be formed from using all of these r relevant variables. Considering this as a 2^r-dimensional feature space, take $\Omega((2^r/\epsilon)log(1/\delta))$ random labeled data points and learn a disjunction over this 2^r-dimensional feature space; since the VC dimension of this set of disjunctions is 2^r, the usual PAC analysis implies this will learn an ϵ-good disjunction with probability $1 - \delta/2$. A union bound implies both stages (finding variables and learning the disjunction) will succeed with probability at least $1 - \delta$. $\qquad\square$

An alternative approach to the second stage in the proof would be to take $\Omega(2^r \log(2^r/\delta))$ random samples, so that with probability at least $1-\delta/2$, we have at least one data point satisfying each of the 2^r possible conjunctions on the relevant variables; then for each of the conjunctions, we check the label of the example that satisfies it, and if that label is positive, we include that conjunction as a term in our DNF, and otherwise we do not include it. This has the property that, altogether, with probability $1 - \delta$, we construct a DNF that has error rate *zero*.

Another family of DNF studied in the literature are those with a sublinear number of terms. Specifically, [Servedio, 2004] proved that the class of $2^{O(\sqrt{\log n})}$-term *monotone* DNF are learnable under the uniform distribution from labeled data alone. As the following theorem states, we can extend this result to include general $2^{O(\sqrt{\log n})}$-term DNF (including non-monotone) given access to our binary pairwise queries.

Theorem 4.17. *Under the uniform distribution, with binary pairwise queries, we can learn any $2^{O(\sqrt{\log n})}$-term DNF (supposing ϵ to be a constant).*

First, we review some known results from [Servedio, 2004]. For any function $g : \{0,1\}^n \to \{-1,+1\}$, define the $g_{i,1}$ and $g_{i,0}$ functions by the property that any x with $x_i = 1$ has $g_{i,1}(x) = g(x)$, and $g_{i,0}(x) = g(y)$, where $y_j = x_j$ for $j \neq i$ and $y_i = 0$. Then define the influence function $I_i(g) = P(g_{i,0}(x) \neq g_{i,1}(x))$. [Servedio, 2004] developed a procedure, FindVariable, which uses a $\mathrm{poly}(n, 1/\gamma, \log(1/\eta))$ number of random labeled samples, labeled according to any monotone DNF g having at most t terms, and with probability $1 - \eta$, returns a set S of variables (indices in $\{1, \ldots, n\}$) such that every $i \notin S$ has $I_i(g) \leq \gamma$ and every $i \in S$ has $I_i(g) \geq \gamma/2$ and the ith variable is contained in some term in g with at most $\log \frac{32tn}{\gamma}$ variables in it.

Furthermore, [Servedio, 2004] showed that, for any t-term DNF f, if we are provided with a set $S_f \subseteq \{1, \ldots, n\}$ such that every $i \notin S_f$ has $I_i(f) \leq \epsilon/4n$, then we can learn f in time polynomial in n, $|S_f|^{O(\log \frac{t}{\epsilon} \log \frac{1}{\epsilon})}$, and $\log(1/\delta)$. In particular, for $|S_f| = O(t \log \frac{tn}{\epsilon})$ and $t = 2^{O(\sqrt{\log n})}$, this is polynomial in n (though not necessarily in ϵ). Given the set S_f, the learning procedure simply estimates the Fourier coefficients for small subsets of S_f.

Proof of Theorem 4.17. To prove Theorem 4.17, we consider the following procedure. First sample m labeled examples $x^{(1)}, \ldots, x^{(m)}$ at random. Then, for each $j \leq m$, define $K_j(\cdot) = K(x^{(j)}, \cdot)$. Now note that, if we define $\varphi_j(y) = (\varphi_{j1}(y), \ldots, \varphi_{jn}(y))$ by $\varphi_{ji}(y) = 2I[y_i = x_i^{(j)}] - 1$, then we can represent $K_j(\cdot) = (K_j'(\varphi_j(\cdot)) + 1)/2$, where K_j' is a monotone DNF (mapping into $\{-1, +1\}$); specifically, the terms in K_j' correspond to the terms in the target satisfied

by $x^{(j)}$, except none of the literals are negated. We then run $\mathrm{FindVariable}$ for each of these K_j', with $\gamma = \epsilon/m$ and $\eta = \delta/2m$. Let S_f denote the union (over $j \leq m$) of the returned sets of variables. It remains only to show this S_f satisfies the requirements for the procedure of [Servedio, 2004], including the size requirement.

Taking $m = \Omega(\frac{ct}{\epsilon} \log \frac{t}{\delta})$, with probability at least $1 - \delta/4$, every term in the target having probability at least $\epsilon/2ct$ will have at least one of the m examples satisfying it. Suppose this event happens. In particular, this means $\mathrm{error}(\max_j K_j) < \epsilon/2c$. Note that

$$I_i(f) = P(f_{i,0}(x) \neq f_{i,1}(x)) \leq 2P(\max_j K_j(x) \neq f(x)) + P((\max_j K_j)_{i,0}(x) \neq (\max_j K_j)_{i,1}(x))$$

$$< \epsilon/c + \sum_j P((K_j')_{i,0}(x) \neq (K_j')_{i,1}(x)) = \epsilon/c + \sum_j I_j(K_j').$$

Thus, by a union bound, with probability $1 - \delta/2$, any variable $i \notin S_f$ has $I_i(f) < \epsilon/c + m\gamma$, and any variable $i \in S_f$ appears in a term in some K_j' of size at most $\log \frac{32tn}{\gamma}$, and therefore also appear in a corresponding term of this size in f. Suppose this happens. Letting $c = 8n$ and $\gamma = \epsilon/8nm$, we have that any $i \notin S_f$ has $I_i(f) < \epsilon/4n$, while any $i \in S_f$ appears in a term of size at most $\log \frac{256tn^2m}{\epsilon} = O(\log \frac{tn \log(1/\delta)}{\epsilon})$. In particular, this implies $|S_f| = O(t \log \frac{tn \log(1/\delta)}{\epsilon})$, and S_f satisfies the requirements of the method of [Servedio, 2004].

Thus, running the procedure from [Servedio, 2004] with confidence parameter $\delta/4$, a union bound implies the total probability of successfully producing an ϵ-good classifier is at least $1 - \delta$. The above process of constructing S_f is clearly polynomial-time. Then, if $t = 2^{O(\sqrt{\log n})}$, the procedure of [Servedio, 2004] runs in time polynomial in n, $\log(1/\delta)$, and $|S_f|^{O(\log(t/\epsilon) \log(1/\epsilon))}$, which is polynomial in n and $\log(1/\delta)$ (though not necessarily in ϵ). $\qquad \square$

4.5 More Powerful Queries

Theorem 4.18. *If we can construct our own feature vectors in addition to getting random data, then under any distribution we can efficiently properly learn DNF using binary-valued queries.*

72

Proof. Suppose we can adaptively construct our own examples. Suppose the target DNF has $T = poly(n)$ terms. Oracle(x, x') gives the number of terms that x and x' have in common. For any x, let x_{-i} be x but with the ith bit flipped. Let $\bar{x}$ be the negative of x.

Below is an algorithm. **Move(**x**,** x'**)** moves x' away from x by one bit, while trying to maintain at least one common term. **LearnTerm(**x**)** returns a term in the target function.

0. **Move**(x, x')

1. $x'' \leftarrow \bar{x}$

2.   **For** $i = 1, 2, ..., n$ s.t. $x_i = x'_i$

3.   **If** (Oracle$(x, x'') \leq$ Oracle(x, x'_{-i}))

4. $x'' \leftarrow x'_{-i}$

5.   Return x''

0. **LearnTerm**(x)

1.   Replicate x to get x'

2.   **While** (Oracle$(x,$ Move$(x, x')) ! = \emptyset$)

3. $x' \leftarrow$ Move(x, x')

4.   Let $I \leftarrow \{i : Oracle(x, x'_{-i}) = \emptyset\}$

5.   Return x_I (i.e. a conjunction with the literals indexed by I, either positive or negative so that x satisfies it)

0. **LearnDNF**

1.   Initialize all-negative DNF $\hat{h}$

2.   Take $M = poly(n) \gg nT$ random examples S

73

3.   **For** each $x \in S$

4.   **If** Oracle(x,x) > 0 (positive example) and $\hat{h}(x)$ = negative

5.   Add term LearnTerm(x) to $\hat{h}$

6.   Return $\hat{h}$ (a DNF with at most T terms, consistent with all M examples)

When we reach x' such that we can't flip any more bits (not already flipped) without making it so they don't satisfy any terms in common anymore, then the bits these two have in common must form a term in the target DNF, so LearnTerm(x) should still find a term in the target DNF.

□

If we can ask about k-tuples of examples (do they all jointly satisfy a term in common?), we have the following result:

Theorem 4.19. *If we can use query sets of arbitrary sizes (instead of just 2 points), then under any distribution we can efficiently properly learn DNF using binary-valued queries from random data.*

Proof. We take any set of examples and ask the oracle the number of terms all examples in the set have in common. Let S be the query set. The idea is to greedily add the examples to S while keeping some terms in common.

Algorithm:

0. Input : dataset D

1. Initialize S to be an empty set

2. **Do**{

3.   **Do**{

4. $r_{\max} \leftarrow 0$

5.   For each example x in the dataset D

6.   add x to the set S

74

7.   query the combined set S, and let $r = Oracle(S)$, $r_{\max} \leftarrow \max\{r_{\max}, r\}$

8.   If $r = 0$, remove x from S, and otherwise leave it in S and remove x from D

9.   } **Until**($r_{max} = 0$)

10. Learn a "most-specific" conjunction from S and add that term to the hypothesis DNF

11. Reset S to empty set

12. }**Until** ($|D| = 0$)

Each time we add a term to the DNF, the examples in S satisfy some term in the target DNF, because we only add each example if by adding it S still has at least one term in common. So the "most-specific" conjunction consistent with S (i.e., the one with most literals in it, still labeling all of S positive) will not misclassify any negative point as positive. Since whenever we add a new term, there were no additional examples in D that could have satisfied a term in common with the examples in S, after adding the term we have removed from D all examples that satisfy the term S has in common. Therefore, the number of terms in our learnt DNF is at most the number of terms T in the true DNF. If the total number of examples is $\gg nT$ (and say T is $poly(n)$), it will get us a DNF that has at most T terms and correctly labels a $poly(n) \gg nT$ sized dataset. Since the training dataset size is much larger than the size of the classifier, by the Occam bound, the learnt DNF will have small generalization error.

□

4.6   Learning DNF with General Queries: Open Questions

- Is it possible to efficiently learn an arbitrary DNF from random data under arbitrary distributions with numerical-valued queries?

- Is it possible to efficiently learn a DNF with $O(1)$ terms from random data under arbitrary distributions with binary-valued queries?

- Is it possible to efficiently learn a monotone DNF from random data under a uniform distribution with numerical-valued queries? If so, what about binary-valued queries?

4.7 Generalizations

4.7.1 Learning Unions of Halfspaces

Several of the above results generalize nicely to the more general problem of learning unions of halfspaces. Specifically, the queries are of the type "do these two examples satisfy a halfspace in common?" or "how many halfspaces do these two examples satisfy in common?" The generalized forms of Theorem 4.19 and Lemma 4.10 follow by the exact same arguments. In each case, the algorithm finds sets of examples that satisfy some halfspace, such that none of the remaining examples satisfy that halfspace, so for each such set we simply find a linear separator to separate those examples from the rest, and take their union to form our final classifier. A sufficiently large (poly(n,$1/\epsilon$)-sized) set suffices to guarantee this works. It is not so clear how to generalize Theorem 4.7, since it is not clear how to use the sets of examples with the common profiles to learn the halfspaces. The generalized version of Theorem 4.6 actually follows from the result below on learning Voronoi diagrams. The generalized version of Theorem 4.18 is simple, since it is even known that labeled data plus membership queries are sufficient.

4.7.2 Learning Voronoi with General Queries

Consider the space of Voronoi diagrams (vector quantizers); specifically, the target function is constant within each cell of the Voronoi diagram, and there are $\text{poly}(n)$ such cells for a given target function. We define a "same cell" query as asking, for a pair of examples x and y, whether x and y occur in the same cell of the target function. With this type of query, we can efficiently properly learn Voronoi partitions from random data, under arbitrary distributions. To prove this, we simply group the examples in a sufficiently large sample into equivalence classes based on

these same-cell queries. For each pair of such equivalence classes, we find a linear separator that separates them. For each test point, we evaluate these linear separators, which thereby associates the test point with one of the equivalence classes from the training data, and we predict as a label for that point the label associated with that equivalence class. If we have a sufficiently large training set, then there is only a small probability the test point gets placed into a different set of points from those in its own cell.

Chapter 5

Bayesian Active Learning with Arbitrary Binary Valued Queries

Abstract

[1]We investigate the minimum expected number of bits sufficient to encode a random variable X while still being able to recover an approximation of X with expected distance from X at most D: that is, the optimal rate at distortion D, in a one-shot coding setting. We find this quantity is related to the entropy of a Voronoi partition of the values of X based on a maximal D-packing.

5.1   Introduction

In this work, we study the fundamental complexity of lossy coding. We are particularly interested in identifying a key quantity that characterizes the expected number of bits (called the *rate*) required to encode a random variable so that we may recover an approximation within expected distance D (called the *distortion*). This topic is a generalization of the well-known analysis of exact coding by Shannon [Shannon, 1948], where it is known that the optimal expected number

[1]Joint work with Jaime Carbonell and Steve Hanneke.

of bits is precisely characterized by the entropy. There are many problems in which exact coding is not practical or not possible, so that lossy coding becomes necessary: particularly for random variables taking values in uncountably infinite spaces. The topic of code lengths for lossy coding is interesting, both for its direct applications to compression, and also as a general setting in which to derive lower bounds for specializations of the setting.

There is much existing work on lossy binary codes. In the present work, we are interested in a "one-shot" analysis of lossy coding [Kieffer, 1993], in which we wish to encode a single random variable, in contrast to the analysis of "asymptotic" source coding [Cover and Thomas, 2006], in which one wishes to simultaneously encode a sequence of random variables. Of particular relevance to the one-shot coding problem is the analysis of *quantization* methods that balance *distortion* with *entropy* [Gersho, 1979, Kieffer, 1993, Zador, 1982]. In particular, it is now well-known that this approach can yield codes that respect a distortion contraint while nearly minimizing the rate, so that there are near-optimal codes of this type [Kieffer, 1993]. Thus, we have an alternative way to think of the optimal rate, in terms of the rate of the best distortion-constrained quantization method. While this is interesting, in that it allows us to restrict our focus in the design of effective coding techniques, it is not as directly helpful if we wish to understand the behavior of the optimal rate itself. That is, since we do not have an explicit description of the optimal quantizer, it may often be difficult to study the behavior of its rate under various interesting conditions. There exist classic results lower bounding the achievable rates, most notably the famous Shannon lower bound [Shannon, 1959], which under certain restrictions on the source and the distortion metric, is known to be fairly tight in the *asymptotic* analysis of source coding [Linder and Zamir, 1994]. However, there are few general results explicitly and tightly characterizing the (non-asymptotic) optimal rates for one-shot coding. In particular, to our knowledge, only a few special-case calculations of the exact value of this optimal rate have been explicitly carried out, such as vectors of independent Bernoulli or Gaussian random variables [Cover and Thomas, 2006].

Below, we discuss a particular distortion-constrained quantizer, based on a Voronoi partition induced by a maximal packing. We are interested in the *entropy* of this quantizer, as a quantity used to characterize the optimal rate for codes of a given distortion. While it is clear that this entropy upper bounds the optimal rate, as this is the case for *any* distortion-constrained quantizer [Kieffer, 1993], the novelty of our analysis lies in noting the remarkable fact that the entropy of any quantizer constructed in this way also *lower bounds* the optimal rate. In particular, this provides a method for approximately calculating the optimal rate without the need to optimize over all possible quantizers. Our result is general, in that it applies to an arbitrary distribution and an arbitrary distortion measure from a general class of finite-dimensional pseudo-metrics. This generality is noteworthy, as it leads to interesting applications in statistical learning theory, which we describe below.

Our analysis is closely related to various notions that arise in the study of ϵ-entropy [Posner and Rodemich, 1971, Posner, Rodemich, and Rumsey, Jr., 1967], in that we are concerned with the entropy of a Voronoi partition induced by an ϵ-cover. The notion of ϵ-entropy has been related to the optimal rates for a given distortion (under a slightly different model than studied here) [Posner and Rodemich, 1971, Posner, Rodemich, and Rumsey, Jr., 1967]. However, there are some important distinctions, perhaps the most significant of which is that calculating the ϵ-entropy requires a prohibitive optimization of the entropy over all ϵ-covers; in contrast, the entropy term in our analysis can be calculated based on *any* maximal ϵ-packing (which is a particular type of ϵ-cover). Maximal ϵ-packings are easy to construct by greedily adding arbitrary new elements to the packing that are ϵ-far from all elements already added; thus, there is always a straightforward algorithmic approach to applying our results.

5.2 Definitions

We suppose $\mathcal{X}^*$ is an arbitrary (nonempty) set, equipped with a separable pseudo-metric $\rho : \mathcal{X}^* \times \mathcal{X}^* \to [0, \infty)$. [2] We suppose $\mathcal{X}^*$ is accompanied by its Borel σ-algebra induced by ρ. There is additionally a (nonempty, measurable) set $\mathcal{X} \subseteq \mathcal{X}^*$, and we denote by $\bar{\rho} = \sup_{h_1, h_2 \in \mathcal{X}} \rho(h_1, h_2)$. Finally, there is a probability measure π with $\pi(\mathcal{X}) = 1$, and an $\mathcal{X}$-valued random variable X with distribution π, referred to here as the "target." As the distribution is essentially arbitrary, the results below will hold for *any* π.

A *code* is a pair of (measurable) functions (ϕ, ψ). The *encoder*, ϕ, maps any element $x \in \mathcal{X}$ to a binary sequence $\phi(x) \in \bigcup_{q=0}^{\infty} \{0, 1\}^q$ (the *codeword*). The *decoder*, ψ, maps any element $c \in \bigcup_{q=0}^{\infty} \{0, 1\}^q$ to an element $\psi(c) \in \mathcal{X}^*$. For any $q \in \{0, 1, \ldots\}$ and $c \in \{0, 1\}^q$, let $|c| = q$ denote the *length* of c. A *prefix-free* code is any code (ϕ, ψ) such that no $x_1, x_2 \in \mathcal{X}$ have $c^{(1)} = \phi(x_1)$ and $c^{(2)} = \phi(x_2)$ with $c^{(1)} \neq c^{(2)}$ but $\forall i \leq |c^{(1)}|, c_i^{(2)} = c_i^{(1)}$: that is, no codeword is a prefix of another (longer) codeword. Let PF denote the set of all prefix-free binary codes.

Here, we consider a setting where the code (ϕ, ψ) may be *lossy*, in the sense that for some values of $x \in \mathcal{X}$, $\rho(\psi(\phi(x)), x) > 0$. Our objective is to design the code to have small expected loss (in the ρ sense), while maintaining as small of an expected codeword length as possible. Formally, we have the following definition, which essentially describes a notion of optimality for a lossy code.

Definition 5.1. *For any* $\mathrm{D} > 0$, *define the optimal* rate *at distortion* D

$$\mathrm{R(D)} = \inf \Big\{ \mathbb{E}\Big[|\phi(X)| \Big] : (\phi, \psi) \in \mathrm{PF} \text{ with}$$
$$\mathbb{E}\Big[\rho\Big(\psi(\phi(X)), X \Big) \Big] \leq \mathrm{D} \Big\},$$

where the random variable in both expectations is $X \sim \pi$.

For our analysis, we will require a notion of dimensionality for the pseudo-metric ρ. For this,

---

[2]The set $\mathcal{X}^*$ will not play any significant role in the analysis, except to allow for improper learning scenarios to be a special case of our setting.

we adopt the well-known *doubling dimension* [Gupta, Krauthgamer, and Lee, 2003].

Definition 5.2. *Define the* doubling dimension d *as the smallest value d such that, for any $x \in \mathcal{X}$, and any $\epsilon > 0$, the size of the minimal $\epsilon/2$-cover of the ϵ-radius ball around x is at most 2^d.*

That is, for any $x \in \mathcal{X}$ and $\epsilon > 0$, there exists a set $\{x_i\}_{i=1}^{2^d}$ of 2^d elements of $\mathcal{X}$ such that

$$\{x' \in \mathcal{X} : \rho(x', x) \leq \epsilon\} \subseteq \bigcup_{i=1}^{2^d} \{x' \in \mathcal{X} : \rho(x', x_i) \leq \epsilon/2\}.$$

Note that, as defined here, d is a constant (i.e., has no dependence on the x or ϵ in its definition). In the analysis below, we will always assume $d < \infty$. The doubling dimension has been studied for a variety of spaces, originally by Gupta, Krauthgamer, & Lee [Gupta, Krauthgamer, and Lee, 2003], and subsequently by many others. In particular, Bshouty, Li, & Long [Bshouty, Li, and Long, 2009] discuss the doubling dimension of spaces $\mathcal{X}$ of binary classifiers, in the context of statistical learning theory.

5.2.1 Definition of Packing Entropy

Our main result concerns the relation between the optimal rate at a given distortion with the entropy of a certain quantizer. We now turn to defining this latter quantity.

Definition 5.3. *For any $\mathrm{D} > 0$, define $\mathcal{Y}(\mathrm{D}) \subseteq \mathcal{X}$ as a maximal D-packing of $\mathcal{X}$. That is, $\forall x_1, x_2 \in \mathcal{Y}(\mathrm{D})$, $\rho(x_1, x_2) \geq \mathrm{D}$, and $\forall x \in \mathcal{X} \setminus \mathcal{Y}(\mathrm{D})$, $\min_{x' \in \mathcal{Y}(\mathrm{D})} \rho(x, x') < \mathrm{D}$.*

For our purposes, if multiple maximal D-packings are possible, we can choose to define $\mathcal{Y}(\mathrm{D})$ arbitrarily from among these; the results below hold for any such choice. Recall that any maximal D-packing of $\mathcal{X}$ is also a D-cover of $\mathcal{X}$, since otherwise we would be able to add to $\mathcal{Y}(\mathrm{D})$ the $x \in \mathcal{X}$ that escapes the cover. That is, $\forall x \in \mathcal{X}$, $\exists y \in \mathcal{Y}(\mathrm{D})$ s.t. $\rho(x, y) < \mathrm{D}$.

Next we define a complexity measure, a type of entropy, which serves as our primary quantity of interest in the analysis of $\mathrm{R}(\mathrm{D})$. It is specified in terms of a partition induced by $\mathcal{Y}(\mathrm{D})$, defined as follows.

82

Definition 5.4. *For any* $D > 0$, *define*

$$\mathcal{Q}(D) = \left\{ \left\{ x \in \mathcal{X} : z = \operatorname*{argmin}_{y \in \mathcal{Y}(D)} \rho(x, y) \right\} : z \in \mathcal{Y}(D) \right\},$$

where we break ties in the argmin *arbitrarily but consistently (e.g., based on a predefined preference ordering of* $\mathcal{Y}(D)$*).*

Definition 5.5. *For any finite (or countable) partition* $\mathcal{S}$ *of* $\mathcal{X}$ *into measurable regions (subsets), define the* entropy *of* $\mathcal{S}$

$$\mathcal{H}(\mathcal{S}) = -\sum_{S \in \mathcal{S}} \pi(S) \log_2 \pi(S).$$

In particular, we will be interested in the quantity $\mathcal{H}(\mathcal{Q}(D))$ in the analysis below.

5.3 Main Result

Our main result can be summarized as follows. Note that, since we took the distribution π to be *arbitrary* in the above definitions, this result holds for *any* given π.

Theorem 5.6. *If* $d < \infty$ *and* $\bar{\rho} < \infty$, *then there is a constant* $c = O(d)$ *such that* $\forall D \in (0, \bar{\rho}/2)$,

$$\mathcal{H}\left(\mathcal{Q}\left(D \log_2(\bar{\rho}/D) \right) \right) - c \leq R(D) \leq \mathcal{H}\left(\mathcal{Q}(D) \right) + 1.$$

It should not be surprising that entropy terms play a key role in this result, as the entropy is essential to the analysis of exact coding [Shannon, 1948]. Furthermore, $R(D)$ is tightly characterized by the minimum achievable entropy among all quantizers of distortion at most D [Kieffer, 1993]. The interesting aspect of Theorem 5.6 is that we can explicitly describe a particular quantizer with near-optimal rate, and its entropy can be explicitly calculated for a variety of scenarios $(\mathcal{X}, \rho, \pi)$. As for the behavior of $R(D)$ within the range between the upper and lower bounds of Theorem 5.6, we should expect the upper bound to be tight when high-probability subsets of the regions in $\mathcal{Q}(D)$ are point-wise well-separated, while $R(D)$ may be much smaller (perhaps closer to the lower bound) when this is violated to a large degree, for reasons described in the proof below.
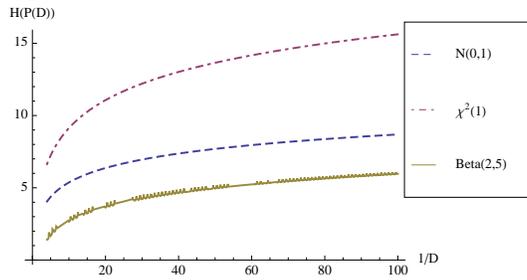
Figure 5.1: Plots of $\mathcal{H}(\mathcal{Q}(D))$ as a function of $1/D$, for various distributions π on $\mathcal{X} = \mathbb{R}$.

Although this result is stated for bounded psuedo-metrics ρ, it also has implications for unbounded ρ. In particular, the proof of the upper bound holds as-is for unbounded ρ. Furthermore, we can always use this lower bound to construct a lower bound for unbounded ρ, simply restricting to a bounded subset of $\mathcal{X}$ with constant probability and calculating the lower bound for that region. For instance, to get a lower bound for π as a Gaussian distribution on $\mathbb{R}$, we might note that $\pi([-1/2, 1/2])$ times the expected loss under the *conditional* $\pi(\cdot|[-1/2, 1/2])$ lower bounds the total expected loss. Thus, calculating the lower bound of Theorem 5.6 under the conditional $\pi(\cdot|[-1/2, 1/2])$ while replacing D with $D/\pi([-1/2, 1/2])$ provides a lower bound on $R(D)$.

To get a feel for the behavior of $\mathcal{H}(\mathcal{Q}(D))$, we have plotted it as a function of $1/D$ for several distributions, in Figure 5.1.

5.4 Proof of Theorem 5.6

We first state a lemma, due to Gupta, Krauthgamer, & Lee [Gupta, Krauthgamer, and Lee, 2003], which will be useful in the proof of Theorem 5.6.

Lemma 5.7. *[Gupta, Krauthgamer, and Lee, 2003] For any $\gamma \in (0, \infty)$, $\delta \in [\gamma, \infty)$, and $x \in \mathcal{X}$,*

$$|\{x' \in \mathcal{Y}(\gamma) : \rho(x', x) \leq \delta\}| \leq \left(\frac{4\delta}{\gamma}\right)^d.$$

In particular, note that this lemma implies that the minimum of $\rho(x, y)$ over $y \in \mathcal{Y}(D)$ is always *achieved* in Definition 5.4, so that $\mathcal{Q}(D)$ is well-defined.

We are now ready for the proof of Theorem 5.6.

Proof of Theorem 5.6. Throughout the proof, we will consider a set-valued random quantity $Q_{\mathrm{D}}(X)$ with value equal to the set in $\mathcal{Q}(\mathrm{D})$ containing X, and a corresponding $\mathcal{X}$-valued random quantity $Y_{\mathrm{D}}(X)$ with value equal the sole point in $Q_{\mathrm{D}}(X) \cap \mathcal{Y}(\mathrm{D})$: that is, the target's nearest representative in the D-packing. Note that, by Lemma 5.7, $|\mathcal{Y}(\mathrm{D})| < \infty$ for all $\mathrm{D} \in (0, 1)$. We will also adopt the usual notation for entropy (e.g., $\mathcal{H}(Q_{\mathrm{D}}(X))$) and conditional entropy (e.g., $\mathcal{H}(Q_{\mathrm{D}}(X)|Z)$) [Cover and Thomas, 2006], both in base 2.

To establish the upper bound, we simply take ϕ as the Huffman code for the random quantity $Q_{\mathrm{D}}(X)$ [Cover and Thomas, 2006, Huffman, 1952]. It is well-known that the expected length of a Huffman code for $Q_{\mathrm{D}}(X)$ is at most $\mathcal{H}(Q_{\mathrm{D}}(X)) + 1$ (in fact, is equal $\mathcal{H}(Q_{\mathrm{D}}(X))$ when the probabilities are powers of 2) [Cover and Thomas, 2006, Huffman, 1952], and each possible value of $Q_{\mathrm{D}}(X)$ is assigned a unique codeword so that we can perfectly recover $Q_{\mathrm{D}}(X)$ (and thus also $Y_{\mathrm{D}}(X)$) based on $\phi(X)$. In particular, define $\psi(\phi(X)) = Y_{\mathrm{D}}(X)$. Finally, recall that any maximal D-packing is also a D-*cover*. Thus, since every element of the set $Q_{\mathrm{D}}(X)$ has $Y_{\mathrm{D}}(X)$ as its closest representative in $\mathcal{Y}(\mathrm{D})$, we must have $\rho(X, \psi(\phi(X))) = \rho(X, Y_{\mathrm{D}}(X)) < \mathrm{D}$. In fact, as this proof never relies on $\bar{\rho} < \infty$, this establishes the upper bound even in the case $\bar{\rho} = \infty$.

The proof of the lower bound is somewhat more involved, though the overall idea is simple enough. Essentially, the lower bound would be straightforward if the regions of $\mathcal{Q}(\mathrm{D} \log_2(\bar{\rho}/\mathrm{D}))$ were separated by some distance, since we could make an argument based on Fano's inequality to say that since any $\hat{X} = \psi(\phi(X))$ is "close" to at most one region, the expected distance from X is at least as large as half this inter-region distance times a quantity proportional to the conditional entropy $\mathcal{H}(Q_{\mathrm{D}}(X)|\phi(X))$, so that $\mathcal{H}(\phi(X))$ can be related to $\mathcal{H}(Q_{\mathrm{D}}(X))$.

However, the general case is not always so simple, as the regions can generally be quite close to each other (even adjacent), so that it is possible for $\hat{X}$ to be close to multiple regions. Thus, the proof will first "color" the regions of $\mathcal{Q}(\mathrm{D} \log_2(\bar{\rho}/\mathrm{D}))$ in a way that guarantees no two regions of the same color are within distance $\mathrm{D} \log_2(\bar{\rho}/\mathrm{D})$ of each other. Then we apply the above simple

argument for each color separately (i.e., lower bounding the expected distance from X under the conditional given the color of $Q_{\mathrm{D}\log_2(\bar{\rho}/\mathrm{D})}(X)$ by a function of the conditional entropy under the conditional), and average over the colors to get a global lower bound. The details follow.

Fix any $\mathrm{D} \in (0, \bar{\rho}/2)$, and for brevity let $\alpha = \mathrm{D}\log_2(\bar{\rho}/\mathrm{D})$. We suppose (ϕ, ψ) is some prefix-free binary code.

Define a function $\mathcal{K} : \mathcal{Q}(\alpha) \to \mathbb{N}$ such that $\forall Q_1, Q_2 \in \mathcal{Q}(\alpha)$,

$$\mathcal{K}(Q_1) = \mathcal{K}(Q_2) \implies \inf_{x_1 \in Q_1, x_2 \in Q_2} \rho(x_1, x_2) \geq \alpha, \tag{5.1}$$

and suppose $\mathcal{K}$ has minimum $\mathcal{H}(\mathcal{K}(Q_\alpha(X)))$ subject to (5.1). We will refer to $\mathcal{K}(Q)$ as the *color* of Q.

Now we are ready to bound the expected distance from X. Let $\hat{X} = \psi(\phi(X))$, and let $Q_\alpha(\hat{X}; \mathcal{K})$ denote the set $Q \in \mathcal{Q}(\alpha)$ having $\mathcal{K}(Q) = \mathcal{K}$ with smallest $\inf_{x \in Q} \rho(x, \hat{X})$ (breaking ties arbitrarily). We know

$$\mathbb{E}[\rho(\hat{X}, X)] = \mathbb{E}\left[\mathbb{E}[\rho(\hat{X}, X)|\mathcal{K}(Q_\alpha(X))]\right]. \tag{5.2}$$

Furthermore, by (5.1) and a triangle inequality, we know no $\hat{X}$ can be closer than $\alpha/2$ to more than one $Q \in \mathcal{Q}(\alpha)$ of a given color. Therefore,

$\mathbb{E}[\rho(\hat{X}, X)|\mathcal{K}(Q_\alpha(X))]$

$$\geq \frac{\alpha}{2}\mathbb{P}(Q_\alpha(\hat{X}; \mathcal{K}(Q_\alpha(X))) \neq Q_\alpha(X)|\mathcal{K}(Q_\alpha(X))). \tag{5.3}$$

By Fano's inequality, we have

$\mathbb{E}\left[\mathbb{P}(Q_\alpha(\hat{X}; \mathcal{K}(Q_\alpha(X))) \neq Q_\alpha(X)|\mathcal{K}(Q_\alpha(X)))\right]$

$$\geq \frac{\mathcal{H}(Q_\alpha(X)|\phi(X), \mathcal{K}(Q_\alpha(X))) - 1}{\log_2 |\mathcal{Y}(\alpha)|}. \tag{5.4}$$

It is generally true that, for a prefix-free binary code $\phi(X)$, $\phi(X)$ is a lossless prefix-free binary code for itself (i.e., with the identity decoder), so that the classic entropy lower bound on average code length [Cover and Thomas, 2006, Shannon, 1948] implies $\mathcal{H}(\phi(X)) \leq \mathbb{E}[|\phi(X)|]$.

86

Also, recalling that $\mathcal{Y}(\alpha)$ is maximal, and therefore also an α-cover, we have that any $Q_1, Q_2 \in \mathcal{Q}(\alpha)$ with $\inf_{x_1 \in Q_1, x_2 \in Q_2} \rho(x_1, x_2) \leq \alpha$ have $\rho(Y_\alpha(x_1), Y_\alpha(x_2)) \leq 3\alpha$ (by a triangle inequality). Therefore, Lemma 5.7 implies that, for any given $Q_1 \in \mathcal{Q}(\alpha)$, there are at most 12^d sets $Q_2 \in \mathcal{Q}(\alpha)$ with $\inf_{x_1 \in Q_1, x_2 \in Q_2} \rho(x_1, x_2) \leq \alpha$. We therefore know there exists a function $\mathcal{K}' : \mathcal{Q}(\alpha) \to \mathbb{N}$ satisfying (5.1) such that $\max_{Q \in \mathcal{Q}(\alpha)} \mathcal{K}'(Q) \leq 12^d$ (i.e., we need at most 12^d colors to satisfy (5.1)). That is, if we consider coloring the sets $Q \in \mathcal{Q}(\alpha)$ sequentially, for any given Q_1 not yet colored, there are $< 12^d$ sets $Q_2 \in \mathcal{Q}(\alpha) \setminus \{Q_1\}$ within α of it, so there must exist a color among $\{1, \ldots, 12^d\}$ not used by any of them, and we can choose that for $\mathcal{K}'(Q_1)$. In particular, by our choice of $\mathcal{K}$ to minimize $\mathcal{H}(\mathcal{K}(Q_\alpha(X)))$ subject to (5.1), this implies

$$\mathcal{H}(\mathcal{K}(Q_\alpha(X))) \leq \mathcal{H}(\mathcal{K}'(Q_\alpha(X))) \leq \log_2(12^d) \leq 4d.$$

Thus,

$$
\begin{aligned}
&\mathcal{H}(Q_\alpha(X)|\phi(X), \mathcal{K}(Q_\alpha(X))) \\
&= \mathcal{H}(Q_\alpha(X), \phi(X), \mathcal{K}(Q_\alpha(X))) \\
&\qquad\qquad - \mathcal{H}(\phi(X)) - \mathcal{H}(\mathcal{K}(Q_\alpha(X))|\phi(X)) \\
&\geq \mathcal{H}(Q_\alpha(X)) - \mathcal{H}(\phi(X)) - \mathcal{H}(\mathcal{K}(Q_\alpha(X))) \\
&\geq \mathcal{H}(Q_\alpha(X)) - \mathbb{E}\left[|\phi(X)|\right] - 4d \\
&= \mathcal{H}(\mathcal{Q}(\alpha)) - \mathbb{E}\left[|\phi(X)|\right] - 4d.
\end{aligned}
\tag{5.5}
$$

Thus, combining (5.2), (5.3), (5.4), and (5.5), we have

$$
\begin{aligned}
\mathbb{E}[\rho(\hat{X}, X)] &\geq \frac{\alpha}{2} \frac{\mathcal{H}(\mathcal{Q}(\alpha)) - \mathbb{E}\left[|\phi(X)|\right] - 4d - 1}{\log_2 |\mathcal{Y}(\alpha)|} \\
&\geq \frac{\alpha}{2} \frac{\mathcal{H}(\mathcal{Q}(\alpha)) - \mathbb{E}\left[|\phi(X)|\right] - 4d - 1}{d \log_2(4\bar{\rho}/\alpha)},
\end{aligned}
$$

where the last inequality follows from Lemma 5.7.

Thus, for any code with

$$\mathbb{E}\left[|\phi(X)|\right] < \mathcal{H}(\mathcal{Q}(\alpha)) - 4d - 1 - 2d \frac{\log_2(4\bar{\rho}/\mathrm{D})}{\log_2(\bar{\rho}/\mathrm{D})},$$

we have $\mathbb{E}[\rho(\hat{X}, X)] > \mathrm{D}$, which implies

$$\mathrm{R(D)} \geq \mathcal{H}(\mathcal{Q}(\alpha)) - 4d - 1 - 2d\frac{\log_2(4\bar{\rho}/\mathrm{D})}{\log_2(\bar{\rho}/\mathrm{D})}.$$

Since $\log_2(4\bar{\rho}/\mathrm{D})/\log_2(\bar{\rho}/\mathrm{D}) \leq 3$, we have

$$\mathrm{R(D)} \geq \mathcal{H}(\mathcal{Q}(\alpha)) - O(d).$$

$\square$

5.5   Application to Bayesian Active Learning

As an example, in the special case of the problem of learning a binary classifier, as studied by
[Haussler, Kearns, and Schapire, 1994a] and [Freund, Seung, Shamir, and Tishby, 1997], $\mathcal{X}^*$ is
the set of all measurable classifiers $h : \mathcal{Z} \to \{-1, +1\}$, $\mathcal{X}$ is called the "concept space," X is
called the "target function," and $\rho(X_1, X_2) = \mathbb{P}(X_1(Z) \neq X_2(Z))$, where Z is some $\mathcal{Z}$-valued
random variable. In particular, $\rho(X_1, X)$ is called the "error rate" of X_1.

We may then discuss a *learning protocol* based on binary-valued queries. That is, we sup-
pose some learning machine is able to pose yes/no questions to an oracle, and based on the
responses it proposes a *hypothesis* $\hat{X}$. We may ask how many such yes/no questions must the
learning machine pose (in expectation) before being able to produce a hypothesis $\hat{X} \in \mathcal{X}^*$ with
$\mathbb{E}[\rho(\hat{X}, X)] \leq \epsilon$, known as the *query complexity*.

If the learning machine is allowed to pose *arbitrary* binary-valued queries, then this setting is
precisely a special case of the general lossy coding problem studied above. That is, any learning
machine that asks a sequence of yes/no questions before terminating and returning some $\hat{X} \in \mathcal{X}^*$
can be thought of as a binary decision tree (no = left, yes = right), with the return $\hat{X}$ values stored
in the leaf nodes. Transforming each root-to-leaf path in the decision tree into a codeword (left
= 0, right = 1), we see that the algorithm corresponds to a prefix-free binary code. Conversely,
given any prefix-free binary code, we can construct an algorithm based on sequentially asking

queries of the form "what is the first bit in the codeword $\phi(X)$ for X?", "what is the second bit in the codeword $\phi(X)$ for X?", etc., until we obtain a complete codeword, at which point we return the value that codeword decodes to. From this perspective, the query complexity is precisely $R(\epsilon)$.

This general problem of learning with arbitrary binary-valued queries was studied previously by Kulkarni, Mitter, & Tsitsiklis [Kulkarni, Mitter, and Tsitsiklis, 1993], in a *minimax* analysis (studying the worst-case value of X). In particular, they find that for a given distribution for Z, the worst-case query complexity is essentially characterized by $\log|\mathcal{Y}(\epsilon)|$. The techniques employed are actually far more general than the classifier-learning problem, and actually apply to any pseudo-metric space. Thus, we can abstractly think of their work as a minimax analysis of lossy coding.

In addition to being quite interesting in their own right, the results of Kulkarni, Mitter, & Tsitsiklis [Kulkarni, Mitter, and Tsitsiklis, 1993] have played a significant role in the recent developments in active learning with *label request* queries for binary classification [Dasgupta, 2005, Hanneke, 2007a,b], in which the learning machine may only ask questions of the form, "What is the value $X(z)$?" for certain values $z \in \mathcal{Z}$. Since label requests can be viewed as a type of binary-valued query, the number of label requests necessary for learning is naturally lower bounded by the number of *arbitrary* binary-valued queries necessary for learning. We therefore always expect to see some term relating to $\log|\mathcal{Y}(\epsilon)|$ in any minimax query complexity results for active learning with label requests (though this factor is typically represented by its upper bound: $\propto V \cdot \log(1/\epsilon)$, where V is the VC dimension).

Similarly to how the work of Kulkarni, Mitter, & Tsitsiklis [Kulkarni, Mitter, and Tsitsiklis, 1993] can be used to argue that $\log|\mathcal{Y}(\epsilon)|$ is a lower bound on the minimax query complexity of active learning with label requests, Theorem 5.6 can be used to argue that $\mathcal{H}(\mathcal{Q}(\epsilon \log_2(1/\epsilon))) - O(d)$ is a lower bound on the query complexity of learning relative to a given distribution for X (called a *prior*, in the language of Bayesian statistics), rather than the worst-case value of X.

Furthermore, as with [Kulkarni, Mitter, and Tsitsiklis, 1993], this lower bound remains valid for learning with label requests, since label requests are a type of binary-valued query. Thus, we should expect a term related to $\mathcal{H}(\mathcal{Q}(\epsilon))$ or $\mathcal{H}(\mathcal{Q}(\epsilon \log_2(1/\epsilon)))$ to appear in any tight analysis of the query complexity of Bayesian learning with label requests.

5.6   Open Problems

In our present context, there are several interesting questions, such as whether the $\log(\bar{\rho}/\mathrm{D})$ factor in the entropy argument of the lower bound can be removed, whether the additive constant in the lower bound might be improved, and in particular whether a similar result might be obtained without assuming $d < \infty$ (e.g., in the statistical learning special case, by making a VC class assumption instead).

Chapter 6

The Sample Complexity of Self-Verifying Bayesian Active Learning

Abstract

[1]We prove that access to a prior distribution over target functions can dramatically improve the sample complexity of self-terminating active learning algorithms, so that it is always better than the known results for prior-dependent passive learning. In particular, this is in stark contrast to the analysis of prior-independent algorithms, where there are simple known learning problems for which no self-terminating algorithm can provide this guarantee for all priors.

6.1   Introduction and Background

Active learning is a powerful form of supervised machine learning characterized by interaction between the learning algorithm and supervisor during the learning process. In this work, we consider a variant known as *pool-based* active learning, in which a learning algorithm is given access to a (typically very large) collection of unlabeled examples, and is able to select any of

---

[1]Joint work with Jaime Carbonell and Steve Hanneke.

those examples, request the supervisor to label it (in agreement with the target concept), then after receiving the label, selects another example from the pool, etc. This sequential label-requesting process continues until some halting criterion is reached, at which point the algorithm outputs a function, and the objective is for this function to closely approximate the (unknown) target concept in the future. The primary motivation behind pool-based active learning is that, often, unlabeled examples are inexpensive and available in abundance, while annotating those examples can be costly or time-consuming; as such, we often wish to select only the informative examples to be labeled, thus reducing information-redundancy to some extent, compared to the baseline of selecting the examples to be labeled uniformly at random from the pool (passive learning).

There has recently been an explosion of fascinating theoretical results on the advantages of this type of active learning, compared to passive learning, in terms of the number of labels required to obtain a prescribed accuracy (called the *sample complexity*): e.g., [Balcan, Broder, and Zhang, 2007a, Balcan, Beygelzimer, and Langford, 2009, Balcan, Hanneke, and Vaughan, 2010, Beygelzimer, Dasgupta, and Langford, 2009, Castro and Nowak, 2008, Dasgupta, 2004, 2005, Dasgupta, Hsu, and Monteleoni, 2007b, Dasgupta, Kalai, and Monteleoni, 2009, Freund, Seung, Shamir, and Tishby, 1997, Friedman, 2009, Hanneke, 2007a,b, 2009, 2011, Kääriäinen, 2006, Koltchinskii, 2010, Nowak, 2008, Wang, 2009]. In particular, [Balcan, Hanneke, and Vaughan, 2010] show that in noise-free binary classifier learning, for any passive learning algorithm for a concept space of finite VC dimension, there exists an active learning algorithm with asymptotically much smaller sample complexity for any nontrivial target concept. In later work, [Hanneke, 2009] strengthens this result by removing a certain strong dependence on the distribution of the data in the learning algorithm. Thus, it appears there are profound advantages to active learning compared to passive learning.

However, the ability to rapidly converge to a good classifier using only a small number of labels is only one desirable quality of a machine learning method, and there are other qualities that may also be important in certain scenarios. In particular, the ability to *verify* the performance

of a learning method is often a crucial part of machine learning applications, as (among other things) it helps us determine whether we have enough data to achieve a desired level of accuracy with the given method. In passive learning, one common practice for this verification is to hold out a random sample of labeled examples as a *validation sample* to evaluate the trained classifier (e.g., to determine when training is complete). It turns out this technique is not feasible in active learning, since in order to be really useful as an indicator of whether we have seen enough labels to guarantee the desired accuracy, the number of labeled examples in the random validation sample would need to be much larger than the number of labels requested by the active learning algorithm itself, thus (to some extent) canceling the savings obtained by performing active rather than passive learning. Another common practice in passive learning is to examine the training error rate of the returned classifier, which can serve as a reasonable indicator of performance (after adjusting for model complexity). However, again this measure of performance is not necessarily reasonable for active learning, since the set of examples the algorithm requests the labels of is typically distributed very differently from the test examples the classifier will be applied to after training.

This reasoning indicates that performance verification is (at best) a far more subtle issue in active learning than in passive learning. Indeed, [Balcan, Hanneke, and Vaughan, 2010] note that although the number of labels required to achieve good accuracy is significantly smaller than passive learning, it is often the case that the number of labels required to *verify* that the accuracy is good is not significantly improved. In particular, this phenomenon can dramatically increase the sample complexity of active learning algorithms that adaptively determine how many labels to request before terminating. In short, if we require the algorithm both to *learn* an accurate concept and to *know* that its concept is accurate, then the number of labels required by active learning is often not significantly smaller than the number required by passive learning.

We should note, however, that the above results were proven for a learning scenario in which the target concept is considered a constant, and no information about the process that generates

this concept is known a priori. Alternatively, we can consider a modification of this problem, so that the target concept can be thought of as a random variable, a sample from a known distribution (called a *prior*) over the space of possible concepts. Such a setting has been studied in detail in the context of passive learning for noise-free binary classification. In particular, [Haussler, Kearns, and Schapire, 1994a] found that for any concept space of finite VC dimension d, for any prior and distribution over data points, $O(d/\varepsilon)$ random labeled examples are sufficient for the expected error rate of the Bayes classifier produced under the posterior distribution to be at most ε. Furthermore, it is easy to construct learning problems for which there is an $\Omega(1/\varepsilon)$ lower bound on the number of random labeled examples required to achieve expected error rate at most ε, by any passive learning algorithm; for instance, the problem of learning threshold classifiers on $[0, 1]$ under a uniform data distribution and uniform prior is one such scenario.

In the context of active learning (again, with access to the prior), [Freund, Seung, Shamir, and Tishby, 1997] analyze the *Query by Committee* algorithm, and find that if a certain information gain quantity for the points requested by the algorithm is lower-bounded by a value g, then the algorithm requires only $O((d/g)\log(1/\varepsilon))$ labels to achieve expected error rate at most ε. In particular, they show that this is satisfied for *constant* g for linear separators under a near-uniform prior, and a near-uniform data distribution over the unit sphere. This represents a marked improvement over the results of [Haussler, Kearns, and Schapire, 1994a] for passive learning, and since the Query by Committee algorithm is self-verifying, this result is highly relevant to the present discussion. However, the condition that the information gains be lower-bounded by a constant is quite restrictive, and many interesting learning problems are precluded by this requirement. Furthermore, there exist learning problems (with finite VC dimension) for which the Query by Committee algorithm makes an expected number of label requests exceeding $\Omega(1/\varepsilon)$. To date, there has not been a general analysis of how the value of g can behave as a function of ε, though such an analysis would likely be quite interesting.

In the present paper, we take a more general approach to the question of active learning with

access to the prior. We are interested in the broad question of whether access to the prior bridges the gap between the sample complexity of *learning* and the sample complexity of learning *with verification*. Specifically, we ask the following question.

Can a prior-dependent self-terminating active learning algorithm for a concept class of finite VC dimension always achieve expected error rate at most ε using $o(1/\varepsilon)$ label requests?

After some basic definitions in Section 6.2, we begin in Section 6.4 with a concrete example, namely interval classifiers under a uniform data density but arbitrary prior, to illustrate the general idea, and convey some of the intuition as to why one might expect a positive answer to this question. In Section 6.5, we present a general proof that the answer is *always* "yes." As the known results for the sample complexity of passive learning with access to the prior are typically $\propto 1/\varepsilon$ [Haussler, Kearns, and Schapire, 1994a], and this is sometimes tight, this represents an improvement over passive learning. The proof is simple and accessible, yet represents an important step in understanding the problem of self-termination in active learning algorithms, and the general issue of the complexity of verification. Also, as this is a result that does *not* generally hold for prior-independent algorithms (even for their "average-case" behavior induced by the prior) for certain concept spaces, this also represents a significant step toward understanding the inherent value of having access to the prior.

6.2 Definitions and Preliminaries

First, we introduce some notation and formal definitions. We denote by $\mathcal{X}$ the *instance space*, representing the range of the unlabeled data points, and we suppose a distribution $\mathcal{D}$ on $\mathcal{X}$, which we will refer to as the *data distribution*. We also suppose the existence of a sequence $X_1, X_2, \ldots$ of i.i.d. random variables, each with distribution $\mathcal{D}$, referred to as the unlabeled data sequence. Though one could potentially analyze the achievable performance as a function of the number of unlabeled points made available to the learning algorithm (cf. [Dasgupta, 2005]), for simplicity in the present work, we will suppose this unlabeled sequence is essentially

inexhaustible, corresponding to the practical fact that unlabeled data are typically available in abundance as they are often relatively inexpensive to obtain. Additionally, there is a set $\mathbb{C}$ of measurable classifiers $h : \mathcal{X} \to \{-1, +1\}$, referred to as the *concept space*. We denote by d the VC dimension of $\mathbb{C}$, and in our present context we will restrict ourselves to spaces $\mathbb{C}$ with $d < \infty$, referred to as a *VC class*. We also have a probability distribution π, called the *prior*, over $\mathbb{C}$, and a random variable $h^* \sim \pi$, called the *target function*; we suppose h^* is independent from the data sequence $X_1, X_2, \ldots$. We adopt the usual notation for conditional expectations and probabilities [Ash and Doléans-Dade, 2000]; for instance, $\mathbb{E}[A|B]$ can be thought of as an expectation of the value A, under the conditional distribution of A given the value of B (which itself is random), and thus the value of $\mathbb{E}[A|B]$ is essentially determined by the value of B. For any measurable $h : \mathcal{X} \to \{-1, +1\}$, define the *error rate* $\mathrm{er}(h) = \mathcal{D}(\{x : h(x) \neq h^*(x)\})$. So far, this setup is essentially identical to that of [Freund, Seung, Shamir, and Tishby, 1997, Haussler, Kearns, and Schapire, 1994a].

The protocol in active learning is the following. An active learning algorithm $\mathcal{A}$ is given as input the prior π, the data distribution $\mathcal{D}$ (though see Section 6.6), and a value $\varepsilon \in (0, 1]$. It also (implicitly) depends on the data sequence $X_1, X_2, \ldots$, and has an indirect dependence on the target function h^* via the following type of interaction. The algorithm may inspect the values X_i for any initial segment of the data sequence, select an index $i \in \mathbb{N}$ to "request" the label of; after selecting such an index, the algorithm receives the value $h^*(X_i)$. The algorithm may then select another index, request the label, receive the value of h^* on that point, etc. This happens for a number of rounds, $N(\mathcal{A}, h^*, \varepsilon, \mathcal{D}, \pi)$, before eventually the algorithm halts and returns a classifier $\hat{h}$. An algorithm is said to be *correct* if $\mathbb{E}\left[\mathrm{er}\left(\hat{h}\right)\right] \leq \varepsilon$ for every $(\varepsilon, \mathcal{D}, \pi)$; that is, given direct access to the prior and the data distribution, and given a specified value ε, a correct algorithm must be guaranteed to have expected error rate at most ε. Define the *expected sample complexity* of $\mathcal{A}$ for $(\mathcal{X}, \mathbb{C}, \mathcal{D}, \pi)$ to be the function $SC(\varepsilon, \mathcal{D}, \pi) = \mathbb{E}[N(\mathcal{A}, h^*, \varepsilon, \mathcal{D}, \pi)]$: the expected number of label requests the algorithm makes.

6.3 Prior-Independent Learning Algorithms

One may initially wonder whether we could achieve this $o(1/\varepsilon)$ result merely by calculating the expected sample complexity of some prior-independent method, thus precluding the need for novel algorithms. Formally, we say an algorithm $\mathcal{A}$ is prior-independent if the conditional distribution of the queries and return value of $\mathcal{A}(\varepsilon, \mathcal{D}, \pi)$ given $\{(X_1, X(X_1)), (X_2, X(X_2)), \ldots\}$ is functionally independent of π. Indeed, for some $\mathbb{C}$ and $\mathcal{D}$, it is known that there *are* prior-independent active learning algorithms $\mathcal{A}$ that have $\mathbb{E}[N(\mathcal{A}, X, \varepsilon, \mathcal{D}, \pi)|X] = o(1/\varepsilon)$ (always); for instance, threshold classifiers have this property under any $\mathcal{D}$, homogeneous linear separators have this property under a uniform $\mathcal{D}$ on the unit sphere in k dimensions, and intervals with positive width on $\mathcal{X} = [0, 1]$ have this property under $\mathcal{D} = \mathrm{Uniform}([0, 1])$ (see e.g., [Dasgupta, 2005]). It is straightforward to show that any such $\mathcal{A}$ will also have $SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi) = o(1/\varepsilon)$ for every π. In particular, the law of total expectation and the dominated convergence theorem imply

$$\lim_{\varepsilon \to 0} \varepsilon SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi) = \lim_{\varepsilon \to 0} \varepsilon \mathbb{E}[\mathbb{E}[N(\mathcal{A}, X, \varepsilon, \mathcal{D}, \pi)|X]]$$
$$= \mathbb{E}\left[\lim_{\varepsilon \to 0} \varepsilon \mathbb{E}[N(\mathcal{A}, X, \varepsilon, \mathcal{D}, \pi)|X]\right] = 0.$$

In these cases, we can think of SC as a kind of *average-case* analysis of these algorithms. However, as we discuss next, there are also many $\mathbb{C}$ and $\mathcal{D}$ for which there is *no* prior-independent algorithm achieving $o(1/\varepsilon)$ sample complexity for *all* priors. Thus, any general result on $o(1/\varepsilon)$ expected sample complexity for π-dependent algorithms would indicate that there is a real advantage to having access to the prior, beyond the apparent *smoothing* effects of an average-case analysis.

As an example of a problem where no prior-independent self-verifying algorithm can achieve $o(1/\varepsilon)$ sample complexity, consider $\mathcal{X} = [0, 1]$, $\mathcal{D} = \mathrm{Uniform}([0, 1])$, and $\mathbb{C}$ as the concept space of *interval classifiers*: $\mathbb{C} = \{\mathbb{I}_{(a,b)}^{\pm} : 0 \le a \le b \le 1\}$, where $\mathbb{I}_{(a,b)}^{\pm}(x) = +1$ if $x \in (a, b)$ and -1 otherwise. Note that because we allow $a = b$, there is a classifier $h_- \in \mathbb{C}$ labeling all of $\mathcal{X}$

negative. For $0 \leq a \leq b \leq 1$, let $\pi_{(a,b)}$ denote the prior with $\pi_{(a,b)}(\{\mathbb{I}^{\pm}_{(a,b)}\}) = 1$. We now show any correct prior-independent algorithm has $\Omega(1/\varepsilon)$ sample complexity for $\pi_{(0,0)}$, following a technique of [Balcan, Hanneke, and Vaughan, 2010]. Consider any $\varepsilon \in (0, 1/144)$ and any prior-independent active learning algorithm $\mathcal{A}$ with $SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi_{(0,0)}) < s = \frac{1}{144\varepsilon}$. Then define $H_{\varepsilon} = \{(12i\varepsilon, 12(i+1)\varepsilon) : i \in \{0, 1, \ldots, \lfloor \frac{1-12\varepsilon}{12\varepsilon} \rfloor \}\}$. Let $\hat{h}_{(a,b)}$ denote the classifier returned by $\mathcal{A}(\varepsilon, \mathcal{D}, \cdot)$ when queries are answered with $X = \mathbb{I}^{\pm}_{(a,b)}$, for $0 \leq a \leq b \leq 1$, and let $R_{(a,b)}$ denote the set of examples (x, y) for which $\mathcal{A}(\varepsilon, \mathcal{D}, \cdot)$ requests labels (including their $y = X(x)$ labels). The point of this construction is that, with such a small number of queries, for many of the $(a, b) \in H_{\varepsilon}$, the algorithm must behave identically for $X = \mathbb{I}^{\pm}_{(a,b)}$ as for $X = \mathbb{I}^{\pm}_{(0,0)}$ (i.e., $R_{(a,b)} = R_{(0,0)}$, and hence $\hat{h}_{(a,b)} = \hat{h}_{(0,0)}$). These $\pi_{(a,b)}$ priors will then witness the fact that $\mathcal{A}$ is not a correct self-verifying algorithm. Formally,

$$
\begin{aligned}
\max_{(a,b) \in H_{\varepsilon}} & \mathbb{E}\left[\mathcal{D}(x : \hat{h}_{(a,b)}(x) \neq \mathbb{I}^{\pm}_{(a,b)}(x))\right] \\
\geq & \frac{1}{|H_{\varepsilon}|} \sum_{(a,b) \in H_{\varepsilon}} \mathbb{E}\left[\mathcal{D}(x : \hat{h}_{(a,b)}(x) \neq \mathbb{I}^{\pm}_{(a,b)}(x))\right] \\
\geq & \frac{1}{|H_{\varepsilon}|} \mathbb{E}\left[\sum_{(a,b) \in H_{\varepsilon} : R_{(a,b)} = R_{(0,0)}} \mathcal{D}(x : \hat{h}_{(a,b)}(x) \neq \mathbb{I}^{\pm}_{(a,b)}(x))\right] \\
\geq & \frac{1}{|H_{\varepsilon}|} \mathbb{E}\left[\sum_{(a,b) \in H_{\varepsilon} : R_{(a,b)} = R_{(0,0)}} \left(12\varepsilon - \min\{\mathcal{D}(x : \hat{h}_{(a,b)}(x) \neq -1), 12\varepsilon\}\right)\right]. \quad (6.1)
\end{aligned}
$$

Since the summation in (6.1) is restricted to (a, b) with $R_{(a,b)} = R_{(0,0)}$, these (a, b) must also have $\hat{h}_{(a,b)} = \hat{h}_{(0,0)}$, so that (6.1) equals

$$
\frac{1}{|H_{\varepsilon}|} \mathbb{E}\left[\sum_{(a,b) \in H_{\varepsilon} : R_{(a,b)} = R_{(0,0)}} \left(12\varepsilon - \min\{\mathcal{D}(x : \hat{h}_{(0,0)}(x) \neq -1), 12\varepsilon\}\right)\right]. \quad (6.2)
$$

Furthermore, for a given $X_1, X_2, \ldots$ sequence, the only $(a, b) \in H_{\varepsilon}$ with $R_{(a,b)} \neq R_{(0,0)}$ are those for which some $(x, -1) \in R_{(0,0)}$ has $x \in (a, b)$; since the $(a, b) \in H_{\varepsilon}$ are disjoint, the

98

above summation has at least $|H_\varepsilon| - |R_{(0,0)}|$ elements in it. Thus, (6.2) is at least

$$\mathbb{E}\left[\left(\frac{|H_\varepsilon| - \min\{|R_{(0,0)}|, |H_\varepsilon|\}}{|H_\varepsilon|}\right)\left(12\varepsilon - \min\{\mathcal{D}(x : \hat{h}_{(0,0)}(x) \neq -1), 12\varepsilon\}\right)\right]$$

$$\geq \mathbb{E}\left[\mathbb{I}\left[|R_{(0,0)}| \leq 3s\right]\mathbb{I}\left[\mathcal{D}(x : \hat{h}_{(0,0)}(x) \neq -1) \leq 6\varepsilon\right]\left(\frac{|H_\varepsilon| - 3s}{|H_\varepsilon|}\right)(12\varepsilon - 6\varepsilon)\right]$$

$$\geq 3\varepsilon\mathbb{P}\left(|R_{(0,0)}| \leq 3s, \mathcal{D}(x : \hat{h}_{(0,0)}(x) \neq -1) \leq 6\varepsilon\right). \qquad (6.3)$$

By Markov's inequality,

$$\mathbb{P}\left(|R_{(0,0)}| > 3s\right) \leq \mathbb{E}[|R_{(0,0)}|]/(3s) = SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi_{(0,0)})/(3s) < 1/3,$$

and $\mathbb{P}\left(\mathcal{D}(x : \hat{h}_{(0,0)}(x) \neq -1) > 6\varepsilon\right) \leq \mathbb{E}\left[\mathcal{D}(x : \hat{h}_{(0,0)}(x) \neq -1)\right]/(6\varepsilon)$, and if $\mathcal{A}$ is a correct self-verifying algorithm, then $\mathbb{E}\left[\mathcal{D}(x : \hat{h}_{(0,0)}(x) \neq -1)\right]/(6\varepsilon) \leq 1/6$. Thus, by a union bound, (6.3) is at least $3\varepsilon(1 - 1/3 - 1/6) = (3/2)\varepsilon > \varepsilon$. Therefore, $\mathcal{A}$ cannot be a correct self-verifying learning algorithm.

6.4 Prior-Dependent Learning: An Example

We begin our exploration of π-dependent active learning with a concrete example, namely interval classifiers under a uniform data density but arbitrary prior, to illustrate how access to the prior can make a difference in the sample complexity. Specifically, consider $\mathcal{X} = [0, 1]$, $\mathcal{D}$ uniform on $[0, 1]$, and the concept space $\mathbb{C}$ of interval classifiers specified in the previous section. For each classifier $h \in \mathbb{C}$, define $w(h) = \mathcal{D}(x : h(x) = +1)$ (the width of the interval h). Note that because we allow $a = b$ in the definition of $\mathbb{C}$, there is a classifier $h_- \in \mathbb{C}$ with $w(h_-) = 0$.

For simplicity, in this example (only) we will suppose the algorithm may request the label of *any* point in $\mathcal{X}$, not just those in the sequence $\{X_i\}$; the same ideas can easily be adapted to the setting where queries are restricted to $\{X_i\}$. Consider an active learning algorithm that sequentially requests the labels $X(x)$ for points x at $1/2$, $1/4$, $3/4$, $1/8$, $3/8$, $5/8$, $7/8$, $1/16$, $3/16$, etc., until (case 1) it encounters an example x with $X(x) = +1$ or until (case 2) the set of

classifiers $V \subseteq \mathbb{C}$ consistent with all observed labels so far satisfies $\mathbb{E}[w(X)|V] \leq \varepsilon$ (which ever comes first). In case 2, the algorithm simply halts and returns the constant classifier that always predicts -1: call it h_-; note that $\mathrm{er}(h_-) = w(X)$. In case 1, the algorithm enters a second phase, in which it performs a binary search (repeatedly querying the midpoint between the closest two -1 and $+1$ points, taking 0 and 1 as known negative points) to the left and right of the observed positive point, halting after $\log_2(4/\varepsilon)$ label requests on each side; this results in estimates of the target's endpoints up to $\pm\varepsilon/4$, so that returning any classifier among the set $V \subseteq \mathbb{C}$ consistent with these labels results in error rate at most ε; in particular, if $\tilde{h}$ is the classifier in V returned, then $\mathbb{E}[\mathrm{er}(\tilde{h})|V] \leq \varepsilon$.

Denoting this algorithm by $\mathcal{A}_{[]}$, and $\hat{h}$ the classifier it returns, we have

$$\mathbb{E}\left[\mathrm{er}\left(\hat{h}\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathrm{er}\left(\hat{h}\right)\Big|V\right]\right] \leq \varepsilon,$$

so that the algorithm is definitely correct.

Note that case 2 will definitely be satisfied after at most $\frac{2}{\varepsilon}$ label requests, and if $w(X) > \varepsilon$, then case 1 will definitely be satisfied after at most $\frac{2}{w(X)}$ label requests, so that the algorithm never makes more than $\frac{2}{\max\{w(X),\varepsilon\}}$ label requests before satisfying one of the two cases. Abbreviating $N(X) = N(\mathcal{A}_{[]}, X, \varepsilon, \mathcal{D}, \pi)$, we have

$$\mathbb{E}\left[N(X)\right]$$

$$= \mathbb{E}\left[N(X)\Big|w(X) = 0\right]\mathbb{P}\left(w(X) = 0\right)$$

$$+ \mathbb{E}\left[N(X)\Big|0 < w(X) \leq \sqrt{\varepsilon}\right]\mathbb{P}\left(0 < w(X) \leq \sqrt{\varepsilon}\right)$$

$$+ \mathbb{E}\left[N(X)\Big|w(X) > \sqrt{\varepsilon}\right]\mathbb{P}\left(w(X) > \sqrt{\varepsilon}\right)$$

$$\leq \mathbb{E}\left[N(X)\Big|w(X) = 0\right]\mathbb{P}\left(w(X) = 0\right) + \frac{2}{\varepsilon}\mathbb{P}\left(0 < w(X) \leq \sqrt{\varepsilon}\right) + \frac{2}{\sqrt{\varepsilon}} + 2\log_2\frac{4}{\varepsilon}. \quad (6.4)$$

The third and fourth terms in (6.4) are $o(1/\varepsilon)$. Since $\mathbb{P}(0 < w(X) \leq \sqrt{\varepsilon}) \to 0$ as $\varepsilon \to 0$, the second term in (6.4) is $o(1/\varepsilon)$ as well. If $\mathbb{P}(w(X) = 0) = 0$, this completes the proof. We focus the rest of the proof on the first term in (6.4), in the case that $\mathbb{P}(w(X) = 0) > 0$: i.e., there is

100

nonzero probability that the target X labels the space all negative. Letting V denote the subset of $\mathbb{C}$ consistent with all requested labels, note that on the event $w(X) = 0$, after n label requests (for $n + 1$ a power of 2) we have $\max_{h \in V} w(h) \leq 1/n$. Thus, for any value $\gamma \in (0, 1)$, after at most $\frac{2}{\gamma}$ label requests, on the event that $w(X) = 0$,

$$\mathbb{E}\left[w(X)\Big|V\right] \leq \frac{\mathbb{E}\left[w(X)\mathbb{I}\left[w(X) \leq \gamma\right]\right]}{\pi(V)} \leq \frac{\mathbb{E}\left[w(X)\mathbb{I}\left[w(X) \leq \gamma\right]\right]}{\mathbb{P}(w(X) = 0)}. \tag{6.5}$$

Now note that, by the dominated convergence theorem,

$$\lim_{\gamma \to 0} \mathbb{E}\left[\frac{w(X)\mathbb{I}\left[w(X) \leq \gamma\right]}{\gamma}\right] = \mathbb{E}\left[\lim_{\gamma \to 0} \frac{w(X)\mathbb{I}\left[w(X) \leq \gamma\right]}{\gamma}\right] = 0.$$

Therefore, $\mathbb{E}\left[w(X)\mathbb{I}\left[w(X) \leq \gamma\right]\right] = o(\gamma)$. If we define γ_ε as the largest value of γ for which $\mathbb{E}\left[w(X)\mathbb{I}\left[w(X) \leq \gamma\right]\right] \leq \varepsilon\mathbb{P}(w(X) = 0)$ (or, say, half the supremum if the maximum is not achieved), then we have $\gamma_\varepsilon \gg \varepsilon$. Combined with (6.5), this implies

$$\mathbb{E}\left[N(X)\Big|w(X) = 0\right] \leq \frac{2}{\gamma_\varepsilon} = o(1/\varepsilon).$$

Thus, all of the terms in (6.4) are $o(1/\varepsilon)$, so that in total $\mathbb{E}[N(X)] = o(1/\varepsilon)$.

In conclusion, for this concept space $\mathbb{C}$ and data distribution $\mathcal{D}$, we have a correct active learning algorithm $\mathcal{A}$ achieving a sample complexity $SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi) = o(1/\varepsilon)$ for all priors π on $\mathbb{C}$.

6.5 A General Result for Self-Verifying Bayesian Active Learning

In this section, we present our main result for improvements achievable by prior-dependent self-verifying active learning: a general result stating that $o(1/\varepsilon)$ expected sample complexity is always achievable for some appropriate prior-dependent active learning algorithm, for *any* $(\mathcal{X}, \mathbb{C}, \mathcal{D}, \pi)$ for which $\mathbb{C}$ has finite VC dimension. Since the known results for the sample complexity of passive learning with access to the prior are typically $\Theta(1/\varepsilon)$ [Haussler, Kearns, and

101

Schapire, 1994a], and since there are known learning problems $(\mathcal{X}, \mathbb{C}, \mathcal{D}, \pi)$ for which every passive learning algorithm requires $\Omega(1/\varepsilon)$ samples, this $o(1/\varepsilon)$ result for active learning represents an improvement over passive learning.

The proof is simple and accessible, yet represents an important step in understanding the problem of self-termination in active learning algorithms, and the general issue of the complexity of verification. Also, since there are problems $(\mathcal{X}, \mathbb{C}, \mathcal{D})$ where $\mathbb{C}$ has finite VC dimension but for which no prior-independent correct active learning algorithm (of the self-terminating type studied here) can achieve $o(1/\varepsilon)$ expected sample complexity for every π, this also represents a significant step toward understanding the inherent value of having access to the prior in active learning.

First, we have a small lemma.

Lemma 6.1. *For any sequence of functions $\phi_n : \mathbb{C} \to [0, \infty)$ such that, $\forall f \in \mathbb{C}$, $\phi_n(f) = o(1/n)$ and $\forall n \in \mathbb{N}$, $\phi_n(f) \leq c/n$ (for an f-independent constant $c \in (0, \infty)$), there exists a sequence $\bar{\phi}_n$ in $[0, \infty)$ such that*

$$\bar{\phi}_n = o(1/n) \quad \text{and} \quad \lim_{n \to \infty} \mathbb{P}\left(\phi_n(X) > \bar{\phi}_n\right) = 0.$$

Proof. For any constant $\gamma \in (0, \infty)$, we have (by Markov's inequality and the dominated convergence theorem)

$$\lim_{n \to \infty} \mathbb{P}\left(n\phi_n(X) > \gamma\right) \leq \frac{1}{\gamma} \lim_{n \to \infty} \mathbb{E}\left[n\phi_n(X)\right]$$
$$= \frac{1}{\gamma} \mathbb{E}\left[\lim_{n \to \infty} n\phi_n(X)\right] = 0.$$

Therefore (by induction), there exists a diverging sequence n_i in $\mathbb{N}$ such that

$$\lim_{i \to \infty} \sup_{n \geq n_i} \mathbb{P}\left(n\phi_n(X) > 2^{-i}\right) = 0.$$

Inverting this, let $i_n = \max\{i \in \mathbb{N} : n_i \leq n\}$, and define $\bar{\phi}_n(X) = (1/n) \cdot 2^{-i_n}$. By construction, $\mathbb{P}\left(\phi_n(X) > \bar{\phi}_n\right) \to 0$. Furthermore, $n_i \to \infty \implies i_n \to \infty$, so that we have

$$\lim_{n \to \infty} n\bar{\phi}_n = \lim_{n \to \infty} 2^{-i_n} = 0,$$

102

implying $\bar{\phi}_n = o(1/n)$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

Theorem 6.2. *For any VC class* $\mathbb{C}$*, there is a correct active learning algorithm* $\mathcal{A}_a$ *that, for every data distribution* $\mathcal{D}$ *and prior* π*, achieves expected sample complexity*

$$SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi) = o(1/\varepsilon).$$

Our approach to proving Theorem 6.2 is via a reduction to established results about (prior-independent) active learning algorithms that are *not* self-verifying. Specifically, consider a slightly different type of active learning algorithm than that defined above: namely, an algorithm $\mathcal{A}_b$ that takes as input a *budget* $n \in \mathbb{N}$ on the number of label requests it is allowed to make, and that after making at most n label requests returns as output a classifier $\hat{h}_n$. Let us refer to any such algorithm as a *budget-based* active learning algorithm. Note that budget-based active learning algorithms are prior-independent (have no direct access to the prior). The following result was proven by [Hanneke, 2009] (see also the related earlier work of [Balcan, Hanneke, and Vaughan, 2010]).

Lemma 6.3. *[Hanneke, 2009] For any VC class* $\mathbb{C}$*, there exists a constant* $c \in (0, \infty)$*, a function* $\mathcal{E}(n; f, \mathcal{D})$*, and a budget-based active learning algorithm* $\mathcal{A}_b$ *such that*

$$\forall \mathcal{D}, \forall f \in \mathbb{C}, \mathcal{E}(n; f, \mathcal{D}) \le c/n \text{ and } \mathcal{E}(n; f, \mathcal{D}) = o(1/n),$$

and $\mathbb{E}\left[\mathrm{er}\left(\mathcal{A}_b(n)\right)\middle|X\right] \le \mathcal{E}(n; X, \mathcal{D})$ *(always).*[2]

That is, equivalently, for any fixed value for the target function, the expected error rate is $o(1/n)$, where the random variable in the expectation is only the data sequence $X_1, X_2, \ldots$. Our task in the proof of Theorem 6.2 is to convert such a budget-based algorithm into one that is correct, self-terminating, and prior-dependent, taking ε as input.

Theorem 6.2. Consider $\mathcal{A}_b$, $\mathcal{E}$, and c as in Lemma 6.3, let $\hat{h}_n$ denote the classifier returned by $\mathcal{A}_b(n)$, and define

$$n_{\pi,\varepsilon} = \min\left\{n \in \mathbb{N} : \mathbb{E}\left[\mathrm{er}\left(\hat{h}_n\right)\right] \le \varepsilon\right\}.$$

---

[2]Furthermore, it is not difficult to see that we can take this $\mathcal{E}$ to be measurable in the X argument.

This value is accessible based purely on access to π and $\mathcal{D}$. Furthermore, we clearly have (by construction) $\mathbb{E}\left[\mathrm{er}\left(\hat{h}_{n_{\pi,\varepsilon}}\right)\right] \leq \varepsilon$. Thus, letting $\mathcal{A}_a$ denote the active learning algorithm taking $(\mathcal{D}, \pi, \varepsilon)$ as input, which runs $\mathcal{A}_b(n_{\pi,\varepsilon})$ and then returns $\hat{h}_{n_{\pi,\varepsilon}}$, we have that $\mathcal{A}_a$ is a *correct* learning algorithm (i.e., its expected error rate is at most ε).

As for the expected sample complexity $SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi)$ achieved by $\mathcal{A}_a$, we have $SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi) \leq n_{\pi,\varepsilon}$, so that it remains only to bound $n_{\pi,\varepsilon}$. By Lemma 6.1, there is a π-dependent function $\mathcal{E}(n; \pi, \mathcal{D})$ such that

$$\pi\left(\{f \in \mathbb{C} : \mathcal{E}(n; f, \mathcal{D}) > \mathcal{E}(n; \pi, \mathcal{D})\}\right) \to 0$$

$$\text{and } \mathcal{E}(n; \pi, \mathcal{D}) = o(1/n).$$

Therefore, by the law of total expectation,

$$\mathbb{E}\left[\mathrm{er}\left(\hat{h}_n\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathrm{er}\left(\hat{h}_n\right)\middle|X\right]\right] \leq \mathbb{E}\left[\mathcal{E}(n; X, \mathcal{D})\right]$$

$$\leq \frac{c}{n}\pi\left(\{f \in \mathbb{C} : \mathcal{E}(n; f, \mathcal{D}) > \mathcal{E}(n; \pi, \mathcal{D})\}\right) + \mathcal{E}(n; \pi, \mathcal{D})$$

$$= o(1/n).$$

If $n_{\pi,\varepsilon} = O(1)$, then clearly $n_{\pi,\varepsilon} = o(1/\varepsilon)$ as needed. Otherwise, since $n_{\pi,\varepsilon}$ is monotonic in ε, we must have $n_{\pi,\varepsilon} \uparrow \infty$ as $\varepsilon \downarrow 0$. In particular, in this latter case we have

$$\lim_{\varepsilon \to 0} \varepsilon \cdot n_{\pi,\varepsilon}$$

$$\leq \lim_{\varepsilon \to 0} \varepsilon \cdot \left(1 + \max\left\{n \geq n_{\pi,\varepsilon} - 1 : \mathbb{E}\left[\mathrm{er}\left(\hat{h}_n\right)\right] > \varepsilon\right\}\right)$$

$$= \lim_{\varepsilon \to 0} \varepsilon \cdot \max_{n \geq n_{\pi,\varepsilon}-1} n\mathbb{I}\left[\mathbb{E}\left[\mathrm{er}\left(\hat{h}_n\right)\right]/\varepsilon > 1\right]$$

$$\leq \lim_{\varepsilon \to 0} \varepsilon \cdot \max_{n \geq n_{\pi,\varepsilon}-1} n\mathbb{E}\left[\mathrm{er}\left(\hat{h}_n\right)\right]/\varepsilon$$

$$= \lim_{\varepsilon \to 0} \max_{n \geq n_{\pi,\varepsilon}-1} n\mathbb{E}\left[\mathrm{er}\left(\hat{h}_n\right)\right] = \limsup_{n \to \infty} n\mathbb{E}\left[\mathrm{er}\left(\hat{h}_n\right)\right] = 0,$$

so that $n_{\pi,\varepsilon} = o(1/\varepsilon)$, as required. $\square$

Theorem 6.2 implies that, if we have *direct* access to the prior distribution of X, regardless of what that prior distribution π is, we can always construct a *self-verifying* active learning algorithm

$\mathcal{A}_a$ that has a guarantee of $\mathbb{E}\left[\mathrm{er}\left(\mathcal{A}_a(\varepsilon, \mathcal{D}, \pi)\right)\right] \leq \varepsilon$ and its expected number of label requests is $o(1/\varepsilon)$. This guarantee is *not* possible for prior-independent self-verifying active learning algorithms.

6.6 Dependence on $\mathcal{D}$ in the Learning Algorithm

The dependence on $\mathcal{D}$ in the algorithm described in the proof of Theorem 6.2 is fairly weak, and we can eliminate any direct dependence on $\mathcal{D}$ by replacing $\mathrm{er}\left(\hat{h}_n\right)$ by a $1-\varepsilon/2$ confidence upper bound based on $M_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$ i.i.d. unlabeled examples $X_1', X_2', \dots, X_{M_\varepsilon}'$ independent from the examples used by the algorithm (e.g., set aside in a pre-processing step, where the bound is calculated via Hoeffding's inequality and a union bound over the values of n that we check, of which there are at most $O(1/\varepsilon)$). Then we simply increase the value of n (starting at some constant, such as 1) until

$$\frac{1}{M_\varepsilon} \sum_{i=1}^{M_\varepsilon} \pi\left(\left\{f \in \mathbb{C} : f\left(X_i'\right) \neq \hat{h}_n\left(X_i'\right)\right\}\right) \leq \varepsilon/2.$$

The expected value of the smallest value of n for which this occurs is $o(1/\varepsilon)$. Note that this only requires access to the prior π, not the data distribution $\mathcal{D}$ (the budget-based algorithm $\mathcal{A}_b$ of [Hanneke, 2009] has no direct dependence on $\mathcal{D}$); if desired for computational efficiency, this dependence may also be estimated by a $1 - \varepsilon/4$ confidence upper bound based on $\Omega\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$ independent samples of X values with distribution π, where for each sample we simulate the execution of $\mathcal{A}_b(n)$ for that (simulated) target function in order to obtain the returned classifier. In particular, note that no actual label requests to the oracle are required during this process of estimating the appropriate label budget $n_{\pi,\varepsilon}$, as all executions of $\mathcal{A}_b$ are *simulated*.

6.7 Inherent Dependence on π in the Sample Complexity

We have shown that for every prior π, the sample complexity is bounded by a $o(1/\varepsilon)$ function. One might wonder whether it is possible that the asymptotic dependence on ε in the sample complexity can be prior-independent, while still being $o(1/\varepsilon)$. That is, we can ask whether there exists a (π-independent) function $s(\varepsilon) = o(1/\varepsilon)$ such that, for every π, there is a correct π-dependent algorithm $\mathcal{A}$ achieving a sample complexity $SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi) = O(s(\varepsilon))$, possibly involving π-dependent constants. Certainly in some cases, such as threshold classifiers, this is true. However, it seems this is not generally the case, and in particular it fails to hold for the space of interval classifiers.

For instance, consider a prior π on the space $\mathbb{C}$ of interval classifiers, constructed as follows. We are given an arbitrary monotonic $g(\varepsilon) = o(1/\varepsilon)$; since $g(\varepsilon) = o(1/\varepsilon)$, there must exist (nonzero) functions $q_1(i)$ and $q_2(i)$ such that $\lim_{i\to\infty} q_1(i) = 0$, $\lim_{i\to\infty} q_2(i) = 0$, and $\forall i \in \mathbb{N}, g(q_1(i)/2^{i+1}) \leq q_2(i) \cdot 2^i$; furthermore, letting $q(i) = \max\{q_1(i), q_2(i)\}$, by monotonicity of g we also have $\forall i \in \mathbb{N}, g(q(i)/2^{i+1}) \leq q(i) \cdot 2^i$, and $\lim_{i\to\infty} q(i) = 0$. Then define a function $p(i)$ with $\sum_{i \in \mathbb{N}} p(i) = 1$ such that $p(i) \geq q(i)$ for infinitely many $i \in \mathbb{N}$; for instance, this can be done inductively as follows. Let $\alpha_0 = 1/2$; for each $i \in \mathbb{N}$, if $q(i) > \alpha_{i-1}$, set $p(i) = 0$ and $\alpha_i = \alpha_{i-1}$; otherwise, set $p(i) = \alpha_{i-1}$ and $\alpha_i = \alpha_{i-1}/2$. Finally, for each $i \in \mathbb{N}$, and each $j \in \{0, 1, \ldots, 2^i - 1\}$, define $\pi\left(\left\{\mathbb{I}^{\pm}_{(j\cdot 2^{-i}, (j+1)\cdot 2^{-i})}\right\}\right) = p(i)/2^i$.

We let $\mathcal{D}$ be uniform on $\mathcal{X} = [0,1]$. Then for each $i \in \mathbb{N}$ s.t. $p(i) \geq q(i)$, there is a $p(i)$ probability the target interval has width 2^{-i}, and given this any algorithm requires $\propto 2^i$ expected number of requests to determine which of these 2^i intervals is the target, failing which the error rate is at least 2^{-i}. In particular, letting $\varepsilon_i = p(i)/2^{i+1}$, any correct algorithm has sample complexity at least $\propto p(i) \cdot 2^i$ for $\varepsilon = \varepsilon_i$. Noting $p(i) \cdot 2^i \geq q(i) \cdot 2^i \geq g(q(i)/2^{i+1}) \geq g(\varepsilon_i)$, this implies there exist arbitrarily small values of $\varepsilon > 0$ for which the optimal sample complexity is at least $\propto g(\varepsilon)$, so that the sample complexity is *not* $o(g(\varepsilon))$.

For any $s(\varepsilon) = o(1/\varepsilon)$, there exists a monotonic $g(\varepsilon) = o(1/\varepsilon)$ such that $s(\varepsilon) = o(g(\varepsilon))$.

Thus, constructing π as above for this g, we have that the sample complexity is not $o(g(\varepsilon))$, and therefore not $O(s(\varepsilon))$. So at least for the space of interval classifiers, the specific $o(1/\varepsilon)$ asymptotic dependence on ε is inherently π-dependent. This argument also illustrates that the $o(1/\varepsilon)$ result in Theorem 6.2 is essentially the strongest possible at this level of generality (i.e., without saying more about $\mathbb{C}$, $\mathcal{D}$, or π).

Chapter 7

Prior Estimation for Transfer Learning

Abstract

[1]We explore a transfer learning setting, in which a finite sequence of target concepts are sampled independently with an unknown distribution from a known family. We study the total number of labeled examples required to learn all targets to an arbitrary specified expected accuracy, focusing on the asymptotics in the number of tasks and the desired accuracy. Our primary interest is formally understanding the fundamental benefits of transfer learning, compared to learning each target independently from the others. Our approach to the transfer problem is general, in the sense that it can be used with a variety of learning protocols.

7.1   Introduction

Transfer learning reuses knowledge from past related tasks to ease the process of learning to perform a new task. The goal of transfer learning is to leverage previous learning and experience to more efficiently learn novel, but related, concepts, compared to what would be possible without this prior experience. The utility of transfer learning is typically measured by a reduction in

---

[1]Joint work with Jaime Carbonell and Steve Hanneke

the number of training examples required to achieve a target performance on a sequence of related learning problems, compared to the number required for unrelated problems: i.e., reduced sample complexity. In many real-life scenarios, just a few training examples of a new concept or process is often sufficient for a human learner to grasp the new concept given knowledge of related ones. For example, learning to drive a van becomes much easier a task if we have already learned how to drive a car. Learning French is somewhat easier if we have already learned English (vs Chinese), and learning Spanish is easier if we know Portuguese (vs German). We are therefore interested in understanding the conditions that enable a learning machine to leverage abstract knowledge obtained as a by-product of learning past concepts, to improve its performance on future learning problems. Furthermore, we are interested in how the magnitude of these improvements grows as the learning system gains more experience from learning multiple related concepts.

The ability to transfer knowledge gained from previous tasks to make it easier to learn a new task can potentially benefit a wide range of real-world applications, including computer vision, natural language processing, cognitive science (e.g., fMRI brain state classification), and speech recognition, to name a few. As an example, consider training a speech recognizer. After training on a number of individuals, a learning system can identify common patterns of speech, such as accents or dialects, each of which requires a slightly different speech recognizer; then, given a new person to train a recognizer for, it can quickly determine the particular dialect from only a few well-chosen examples, and use the previously-learned recognizer for that particular dialect. In this case, we can think of the transferred knowledge as consisting of the common aspects of each recognizer variant and more generally the *distribution* of speech patterns existing in the population these subjects are from. This same type of distribution-related knowledge transfer can be helpful in a host of applications, including all those mentioned above.

Supposing these target concepts (e.g., speech patterns) are sampled independently from a fixed population, having knowledge of the distribution of concepts in the population may often

109

be quite valuable. More generally, we may consider a general scenario in which the target concepts are sampled i.i.d. according to a fixed distribution. As we show below, the number of labeled examples required to learn a target concept sampled according to this distribution may be dramatically reduced if we have direct knowledge of the distribution. However, since in many real-world learning scenarios, we do not have direct access to this distribution, it is desirable to be able to somehow *learn* the distribution, based on observations from a sequence of learning problems with target concepts sampled according to that distribution. The hope is that an estimate of the distribution so-obtained might be almost as useful as direct access to the true distribution in reducing the number of labeled examples required to learn subsequent target concepts. The focus of this paper is an approach to transfer learning based on estimating the distribution of the target concepts. Whereas we acknowledge that there are other important challenges in transfer learning, such as exploring improvements obtainable from transfer under various alternative notions of task relatedness [Ben-David and Schuller, 2003, Evgeniou and Pontil, 2004], or alternative reuses of knowledge obtained from previous tasks [Thrun, 1996], we believe that learning the distribution of target concepts is a central and crucial component in many transfer learning scenarios, and can reduce the total sample complexity across tasks.

Note that it is not immediately obvious that the distribution of targets can even be learned in this context, since we do not have direct access to the target concepts sampled according to it, but rather have only indirect access via a finite number of labeled examples for each task; a significant part of the present work focuses on establishing that as long as these finite labeled samples are larger than a certain size, they hold sufficient information about the distribution over concepts for estimation to be possible. In particular, in contrast to standard results on consistent density estimation, our estimators are not directly based on the target concepts, but rather are only indirectly dependent on these via the labels of a finite number of data points from each task. One desideratum we pay particular attention to is minimizing the number of *extra* labeled examples needed for each task, beyond what is needed for learning that particular target, so that

the benefits of transfer learning are obtained almost as a *by-product* of learning the targets. Our technique is general, in that it applies to any concept space with finite VC dimension; also, the process of learning the target concepts is (in some sense) decoupled from the mechanism of learning the concept distribution, so that we may apply our technique to a variety of learning protocols, including passive supervised learning, active supervised learning, semi-supervised learning, and learning with certain general data-dependent forms of interaction [Hanneke, 2009]. For simplicity, we choose to formulate our transfer learning algorithms in the language of active learning; as we show, this problem can benefit significantly from transfer. Formulations for other learning protocols would follow along similar lines, with analogous theorems; these results are particularly interested when composed with the results on prior-dependent active learning from the previous chapter.

Transfer learning is related at least in spirit to much earlier work on case-based and analogical learning [Carbonell, 1983, 1986, Kolodner (Ed), 1993, Thrun, 1996, Veloso and Carbonell, 1993], although that body of work predated modern machine learning, and focused on symbolic reuse of past problem solving solutions rather than on current machine learning problems such as classification, regression or structured learning. More recently, transfer learning (and the closely related problem of *multitask* learning) has been studied in specific cases with interesting (though sometimes heuristic) approaches [Baxter, 1997, Ben-David and Schuller, 2003, Caruana, 1997, Micchelli and Pontil, 2004, Silver, 2000]. This paper considers a general theoretical framework for transfer learning, based on an Empirical Bayes perspective, and derives rigorous theoretical results on the benefits of transfer. We discuss the relation of this analysis to existing theoretical work on transfer learning below.

7.1.1 Outline of the paper

The remainder of the paper is organized as follows. In Section 7.2 we introduce basic notation used throughout, and survey some related work from the existing literature. In Section 7.3, we

describe and analyze our proposed method for estimating the distribution of target concepts, the key ingredient in our approach to transfer learning, which we then present in Section 7.4.

7.2 Definitions and Related Work

First, we state a few basic notational conventions. We denote $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For any random variable X, we generally denote by $\mathbb{P}_X$ the distribution of X (the induced probability measure on the range of X), and by $\mathbb{P}_{X|Y}$ the regular conditional distribution of X given Y. For any pair of probability measures μ_1, μ_2 on a measurable space $(\Omega, \mathcal{F})$, we define

$$\|\mu_1 - \mu_2\| = \sup_{A \in \mathcal{F}} |\mu_1(A) - \mu_2(A)|.$$

Next we define the particular objects of interest to our present discussion. Let Θ be an arbitrary set (called the *parameter space*), $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a Borel space [Schervish, 1995] (where $\mathcal{X}$ is called the *instance space*), and $\mathcal{D}$ be a fixed distribution on $\mathcal{X}$ (called the *data distribution*). For instance, Θ could be $\mathbb{R}^n$ and $\mathcal{X}$ could be $\mathbb{R}^m$, for some $n, m \in \mathbb{N}$, though more general scenarios are certainly possible as well, including infinite-dimensional parameter spaces. Let $\mathbb{C}$ be a set of measurable classifiers $h : \mathcal{X} \to \{-1, +1\}$ (called the *concept space*), and suppose $\mathbb{C}$ has VC dimension $d < \infty$ [Vapnik, 1982] (such a space is called a *VC class*). $\mathbb{C}$ is equipped with its Borel σ-algebra $\mathcal{B}$, induced by the pseudo-metric $\rho(h, g) = \mathcal{D}(\{x \in \mathcal{X} : h(x) \neq g(x)\})$. Though all of our results can be formulated for general $\mathcal{D}$ in slightly more complex terms, for simplicity throughout the discussion below we suppose ρ is actually a *metric*, in that any $h, g \in \mathbb{C}$ with $h \neq g$ have $\rho(h, g) > 0$; this amounts to a topological assumption on $\mathbb{C}$ relative to $\mathcal{D}$.

For each $\theta \in \Theta$, π_θ is a distribution on $\mathbb{C}$ (called a *prior*). Our only (rather mild) assumption on this family of prior distributions is that $\{\pi_\theta : \theta \in \Theta\}$ be totally bounded, in the sense that $\forall \varepsilon > 0$, $\exists$ *finite* $\Theta_\varepsilon \subseteq \Theta$ s.t. $\forall \theta \in \Theta, \exists \theta_\varepsilon \in \Theta_\varepsilon$ with $\|\pi_\theta - \pi_{\theta_\varepsilon}\| < \varepsilon$. See [Devroye and Lugosi, 2001] for examples of categories of classes that satisfy this.

The general setup for the learning problem is that we have a *true* parameter value $\theta_\star \in \Theta$, and

112

a collection of $\mathbb{C}$-valued random variables $\{h_{t\theta}^*\}_{t\in\mathbb{N},\theta\in\Theta}$, where for a fixed $\theta \in \Theta$ the $\{h_{t\theta}^*\}_{t\in\mathbb{N}}$ variables are i.i.d. with distribution π_θ.

The learning problem is the following. For each $\theta \in \Theta$, there is a sequence

$$\mathcal{Z}_t(\theta) = \{(X_{t1}, Y_{t1}(\theta)), (X_{t2}, Y_{t2}(\theta)), \ldots\},$$

where $\{X_{ti}\}_{t,i\in\mathbb{N}}$ are i.i.d. $\mathcal{D}$, and for each $t, i \in \mathbb{N}$, $Y_{ti}(\theta) = h_{t\theta}^*(X_{ti})$. For $k \in \mathbb{N}$ we denote by $\mathcal{Z}_{tk}(\theta) = \{(X_{t1}, Y_{t1}(\theta)), \ldots, (X_{tk}, Y_{tk}(\theta))\}$. Since the $Y_{ti}(\theta)$ are the actual $h_{t\theta}^*(X_{ti})$ values, we are studying the non-noisy, or *realizable-case*, setting.

The algorithm receives values ε and T as input, and for each $t \in \{1, 2, \ldots, T\}$ in increasing order, it observes the sequence $X_{t1}, X_{t2}, \ldots$, and may then select an index i_1, receive label $Y_{ti_1}(\theta_\star)$, select another index i_2, receive label $Y_{ti_2}(\theta_\star)$, etc. The algorithm proceeds in this fashion, sequentially requesting labels, until eventually it produces a classifier $\hat{h}_t$. It then increments t and repeats this process until it produces a sequence $\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_T$, at which time it halts. To be called *correct*, the algorithm must have a guarantee that $\forall \theta_\star \in \Theta, \forall t \leq T, \mathbb{E}\left[\rho\left(\hat{h}_t, h_{t\theta_\star}^*\right)\right] \leq \varepsilon$, for any values of $T \in \mathbb{N}$ and $\varepsilon > 0$ given as input. We will be interested in the expected number of label requests necessary for a correct learning algorithm, averaged over the T tasks, and in particular in how shared information between tasks can help to reduce this quantity when direct access to $\theta_\star$ is not available to the algorithm.

7.2.1 Relation to Existing Theoretical Work on Transfer Learning

Although we know of no existing work on the theoretical advantages of transfer learning for active learning, the existing literature contains several analyses of the advantages of transfer learning for passive learning. In his classic work, Baxter ([Baxter, 1997] section 4) explores a similar setup for a general form of passive learning, except in a *full* Bayesian setting (in contrast to our setting, often referred to as "empirical Bayes," which includes a constant parameter $\theta_\star$ to be estimated from data). Essentially, [Baxter, 1997] sets up a hierarchical Bayesian model, in which (in our notation) $\theta_\star$ is a random variable with known distribution (hyper-prior), but otherwise the

specialization of Baxter's setting to the pattern recognition problem is essentially identical to our setup above. This hyper-prior does make the problem slightly easier, but generally the results of [Baxter, 1997] are of a different nature than our objectives here. Specifically, Baxter's results on learning from labeled examples can be interpreted as indicating that transfer learning can improve certain *constant factors* in the asymptotic rate of convergence of the average of expected error rates across the learning problems. That is, certain constant complexity terms (for instance, related to the concept space) can be reduced to (potentially much smaller) values related to $\pi_{\theta_\star}$ by transfer learning. Baxter argues that, as the number of tasks grows large, this effectively achieves close to the known results on the sample complexity of passive learning with direct access to $\theta_\star$. A similar claim is discussed by Ando and Zhang [Ando and Zhang, 2004] (though in less detail) for a setting closer to that studied here, where $\theta_\star$ is an unknown parameter to be estimated.

There are also several results on transfer learning of a slightly different variety, in which, rather than having a prior distribution for the target concept, the learner initially has several potential concept spaces to choose from, and the role of transfer is to help the learner select from among these concept spaces [Ando and Zhang, 2005, Baxter, 2000]. In this case, the idea is that one of these concept spaces has the best average minimum achievable error rate per learning problem, and the objective of transfer learning is to perform nearly as well as if we knew which of the spaces has this property. In particular, if we assume the target functions for each task all reside in one of the concept spaces, then the objective of transfer learning is to perform nearly as well as if we knew which of the spaces contains the targets. Thus, transfer learning results in a sample complexity related to the number of learning problems, a complexity term for this best concept space, and a complexity term related to the diversity of concept spaces we have to choose from. In particular, as with [Baxter, 1997], these results can typically be interpreted as giving constant factor improvements from transfer in a passive learning context, at best reducing the complexity constants, from those for the union over the given concept spaces, down to the complexity constants of the single best concept space.

114

In addition to the above works, there are several analyses of transfer learning and multitask learning of an entirely different nature than our present discussion, in that the objectives of the analysis are somewhat different. Specifically, there is a branch of the literature concerned with task *relatedness*, not in terms of the underlying process that generates the target concepts, but rather directly in terms of relations between the target concepts themselves. In this sense, several tasks with related target concepts should be much easier to learn than tasks with unrelated target concepts. This is studied in the context of kernel methods by [Evgeniou and Pontil, 2004, Evgeniou, Micchelli, and Pontil, 2005, Micchelli and Pontil, 2004], and in a more general theoretical framework by [Ben-David and Schuller, 2003]. As mentioned, our approach to transfer learning is based on the idea of estimating the distribution of target concepts. As such, though interesting and important, these notions of direct relatedness of target concepts are not as relevant to our present discussion.

As with [Baxter, 1997], the present work is interested in showing that as the number of tasks grows large, we can effectively achieve a sample complexity close to that achievable with direct access to $\theta_\star$. However, in contrast, we are interested in a general approach to transfer learning and the analysis thereof, leading to concrete results for a variety of learning protocols such as active learning and semi-supervised learning. In particular, our analysis of active learning reveals the interesting phenomenon that transfer learning can sometimes improve the asymptotic dependence on ε, rather than merely the constant factors as in the analysis of [Baxter, 1997].

Our work contrasts with [Baxter, 1997] in another important respect, which significantly changes the way we approach the problem. Specifically, in Baxter's analysis, the results (e.g., [Baxter, 1997] Theorems 4, 6) regard the average loss over the tasks, and are stated as a function of the number of samples per task. This number of samples plays a dual role in Baxter's analysis, since these samples are used both by the individual learning algorithm for each task, and also for the global transfer learning process that provides the learners with information about $\theta_\star$. Baxter is then naturally interested in the rates at which these losses shrink as the sample sizes grow

large, and therefore formulates the results in terms of the asymptotic behavior as the per-task sample sizes grow large. In particular, the results of [Baxter, 1997] involve residual terms which become negligible for large sample sizes, but may be more significant for smaller sample sizes.

In our work, we are interested in decoupling these two roles for the sample sizes; in particular, our results regard only the number of tasks as an asymptotic variable, while the number of samples per task remains bounded. First, we note a very practical motivation for this: namely, non-altruistic learners. In many settings where transfer learning may be useful, it is desirable that the number of labeled examples we need to collect from each particular learning problem never be significantly larger than the number of such examples required to solve that particular problem (i.e., to learn that target concept to the desired accuracy). For instance, this is the case when the learning problems are not all solved by the same individual (or company, etc.), but rather a coalition of cooperating individuals (e.g., hospitals sharing data on clinical trials); each individual may be willing to share the data they used to learn their particular concept, in the interest of making others' learning problems easier; however, they may not be willing to collect significantly *more* data than they themselves need for their own learning problem. We should therefore be particularly interested in studying transfer as a *by-product* of the usual learning process; failing this, we are interested in the minimum possible number of *extra* labeled examples per task to gain the benefits of transfer learning.

The issue of non-altruistic learners also presents a further technical problem in that the individuals solving each task may be unwilling to alter their *method* of gathering data to be more informative for the transfer learning process. That is, we expect the learning process for each task is designed with the sole intention of estimating the target concept, without regard for the global transfer learning problem. To account for this, we model the transfer learning problem in a reduction-style framework, in which we suppose there is some black-box learning algorithm to be run for each task, which takes a prior as input and has a theoretical guarantee of good performance provided the prior is correct. We place almost no restrictions whatsoever on this learning

116

algorithm, including the manner in which it accesses the data. This allows remarkable generality, since this procedure could be passive, active, semi-supervised, or some other kind of query-based strategy. However, because of this generality, we have no guarantee on the information about $\theta_\star$ reflected in the data used by this algorithm (especially if it is an active learning algorithm). As such, we choose not to use the label information gathered by the learning algorithm for each task when estimating the $\theta_\star$, but instead take a small number of *additional* random labeled examples from each task with which to estimate $\theta_\star$. Again, we want to minimize this number of additional samples per task; indeed, in this work we are able to make due with a mere *constant* number of additional samples per task. To our knowledge, no result of this type (estimating $\theta_\star$ using a bounded sample size per learning problem) has previously been established at the level of generality studied here.

7.3 Estimating the Prior

The advantage of transfer learning in this setting is that each learning problem provides some information about $\theta_\star$, so that after solving several of the learning problems, we might hope to be able to *estimate* $\theta_\star$. Then, with this estimate in hand, we can use the corresponding estimated prior distribution in the learning algorithm for subsequent learning problems, to help inform the learning process similarly to how direct knowledge of $\theta_\star$ might be helpful. However, the difficulty in approaching this is how to define such an estimator. Since we do not have direct access to the h_t^* values, but rather only indirect observations via a finite number of example labels, the standard results for density estimation from i.i.d. samples cannot be applied.

The idea we pursue below is to consider the distributions on $\mathcal{Z}_{tk}(\theta_\star)$. These variables *are* directly observable, by requesting the labels of those examples. Thus, for any finite $k \in \mathbb{N}$, this distribution *is* estimable from observable data. That is, using the i.i.d. values $\mathcal{Z}_{1k}(\theta_\star), \ldots, \mathcal{Z}_{tk}(\theta_\star)$, we can apply standard techniques for density estimation to arrive at an estimator of $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$. Then the question is whether the distribution $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$ uniquely characterizes the prior distribution $\pi_{\theta_\star}$:

that is, whether $\pi_{\theta_\star}$ is *identifiable* from $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$.

As an example, consider the space of *half-open interval* classifiers on $[0,1]$: $\mathbb{C} = \{\mathbb{1}^\pm_{[a,b)} :$ $0 \le a \le b \le 1\}$, where $\mathbb{1}^\pm_{[a,b)}(x) = +1$ if $a \le x < b$ and -1 otherwise. In this case, $\pi_{\theta_\star}$ is *not* necessarily identifiable from $\mathbb{P}_{\mathcal{Z}_{t1}(\theta_\star)}$; for instance, the distributions π_{θ_1} and π_{θ_2} characterized by $\pi_{\theta_1}(\{\mathbb{1}^\pm_{[0,1)}\}) = \pi_{\theta_1}(\{\mathbb{1}^\pm_{[0,0)}\}) = 1/2$ and $\pi_{\theta_2}(\{\mathbb{1}^\pm_{[0,1/2)}\}) = \pi_{\theta_2}(\{\mathbb{1}^\pm_{[1/2,1)}\}) = 1/2$ are not distinguished by these one-dimensional distributions. However, it turns out that for this half-open intervals problem, $\pi_{\theta_\star}$ *is* uniquely identifiable from $\mathbb{P}_{\mathcal{Z}_{t2}(\theta_\star)}$; for instance, in the θ_1 vs θ_2 scenario, the conditional probability $\mathbb{P}_{(Y_{t1}(\theta_i),Y_{t2}(\theta_i))|(X_{t1},X_{t2})}((+1,+1)|(1/4,3/4))$ will distinguish π_{θ_1} from π_{θ_2}, and this can be calculated from $\mathbb{P}_{\mathcal{Z}_{t2}(\theta_i)}$. The crucial element of the analysis below is determining the appropriate value of k to uniquely identify $\pi_{\theta_\star}$ from $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$ *in general*. As we will see, $k = d$ (the VC dimension) is *always* sufficient, a key insight for the results that follow. We will also see this is *not* the case for any $k < d$.

To be specific, in order to transfer knowledge from one task to the next, we use a few labeled data points from each task to gain information about $\theta_\star$. For this, for each task t, we simply take the first d data points in the $\mathcal{Z}_t(\theta_\star)$ sequence. That is, we request the labels

$$Y_{t1}(\theta_\star), Y_{t2}(\theta_\star), \ldots, Y_{td}(\theta_\star)$$

and use the points $\mathcal{Z}_{td}(\theta_\star)$ to update an estimate of $\theta_\star$.

The following result shows that this technique does provide a consistent estimator of $\pi_{\theta_\star}$. Again, note that this result is not a straightforward application of the standard approach to consistent estimation, since the observations here are not the $h^*_{t\theta_\star}$ variables themselves, but rather a number of the $Y_{ti}(\theta_\star)$ values. The key insight in this result is that $\pi_{\theta_\star}$ is *uniquely identified* by the joint distribution $\mathbb{P}_{\mathcal{Z}_{td}(\theta_\star)}$ over the first d labeled examples; later, we prove this is *not* necessarily true for $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$ for values $k < d$. This identifiability result is stated below in Corollary 7.6; as we discuss in Section 7.3.1, there is a fairly simple direct proof of this result. However, for our purposes, we will actually require the stronger condition that any $\theta \in \Theta$ with small $\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}\|$ also has small $\|\pi_\theta - \pi_{\theta_\star}\|$. This stronger requirement adds to the complexity

of the proofs. The results in this section are purely concerned with relating distances in the space of $\mathbb{P}_{\mathcal{Z}_{td}(\theta)}$ distributions to the corresponding distances in the space of π_θ distributions; as such, they are not specific to active learning or other learning protocols, and hence are of independent interest.

Theorem 7.1. *There exists an estimator $\hat{\theta}_{T\theta_\star} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star))$, and functions $R :$ $\mathbb{N}_0 \times (0,1] \to [0,\infty)$ and $\delta : \mathbb{N}_0 \times (0,1] \to [0,1]$, such that for any $\alpha > 0$, $\lim\limits_{T\to\infty} R(T,\alpha) = \lim\limits_{T\to\infty} \delta(T,\alpha) = 0$ and for any $T \in \mathbb{N}_0$ and $\theta_\star \in \Theta$,*

$$\mathbb{P}\left(\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\| > R(T,\alpha)\right) \le \delta(T,\alpha) \le \alpha.$$

One important detail to note, for our purposes, is that $R(T,\alpha)$ is independent from $\theta_\star$, so that the value of $R(T,\alpha)$ can be calculated and used within a learning algorithm. The proof of Theorem 7.1 will be established via the following sequence of lemmas. Lemma 7.2 relates distances in the space of priors to distances in the space of distributions on the full data sets. In turn, Lemma 7.3 relates these distances to distances in the space of distributions on a finite number of examples from the data sets. Lemma 7.4 then relates the distances between distributions on any finite number of examples to distances between distributions on d examples. Finally, Lemma 7.5 presents a standard result on the existence of a converging estimator, in this case for the distribution on d examples, for totally bounded families of distributions. Tracing these relations back, they relate convergence of the estimator for the distribution of d examples to convergence of the corresponding estimator for the prior itself.

Lemma 7.2. *For any $\theta, \theta' \in \Theta$ and $t \in \mathbb{N}$,*

$$\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\|.$$

Proof. Fix $\theta, \theta' \in \Theta$, $t \in \mathbb{N}$. Let $\mathbb{X} = \{X_{t1}, X_{t2}, \ldots\}$, $\mathbb{Y}(\theta) = \{Y_{t1}(\theta), Y_{t2}(\theta), \ldots\}$, and for $k \in \mathbb{N}$ let $\mathbb{X}_k = \{X_{t1}, \ldots, X_{tk}\}$. and $\mathbb{Y}_k(\theta) = \{Y_{t1}(\theta), \ldots, Y_{tk}(\theta)\}$. For $h \in \mathbb{C}$, let $c_\mathbb{X}(h) = \{(X_{t1}, h(X_{t1})), (X_{t2}, h(X_{t2})), \ldots\}$.

For $h, g \in \mathbb{C}$, define $\rho_\mathbb{X}(h,g) = \lim\limits_{m\to\infty} \frac{1}{m} \sum_{i=1}^m \mathbb{1}[h(X_{ti}) \ne g(X_{ti})]$ (if the limit exists), and $\rho_{\mathbb{X}_k}(h,g) = \frac{1}{k} \sum_{i=1}^k \mathbb{1}[h(X_{ti}) \ne g(X_{ti})]$. Note that since $\mathbb{C}$ has finite VC dimension, so does

119

the collection of sets $\{\{x : h(x) \neq g(x)\} : h, g \in \mathbb{C}\}$, so that the uniform strong law of large numbers implies that with probability one, $\forall h, g \in \mathbb{C}$, $\rho_{\mathbb{X}}(h, g)$ exists and has $\rho_{\mathbb{X}}(h, g) = \rho(h, g)$ [Vapnik, 1982].

Consider any $\theta, \theta' \in \Theta$, and any $A \in \mathcal{B}$. Then since $\mathcal{B}$ is the Borel σ-algebra induced by ρ, any $h \notin A$ has $\forall g \in A$, $\rho(h, g) > 0$. Thus, if $\rho_{\mathbb{X}}(h, g) = \rho(h, g)$ for all $h, g \in \mathbb{C}$, then $\forall h \notin A$,

$$\forall g \in A, \rho_{\mathbb{X}}(h, g) = \rho(h, g) > 0 \implies \forall g \in A, c_{\mathbb{X}}(h) \neq c_{\mathbb{X}}(g) \implies c_{\mathbb{X}}(h) \notin c_{\mathbb{X}}(A).$$

This implies $c_{\mathbb{X}}^{-1}(c_{\mathbb{X}}(A)) = A$. Under these conditions,

$$\mathbb{P}_{\mathcal{Z}_t(\theta)|\mathbb{X}}(c_{\mathbb{X}}(A)) = \pi_\theta(c_{\mathbb{X}}^{-1}(c_{\mathbb{X}}(A))) = \pi_\theta(A),$$

and similarly for θ'.

Any measurable set C for the range of $\mathcal{Z}_t(\theta)$ can be expressed as $C = \{c_{\bar{x}}(h) : (h, \bar{x}) \in C'\}$ for some appropriate $C' \in \mathcal{B} \otimes \mathcal{B}_{\mathcal{X}}^\infty$. Letting $C'_{\bar{x}} = \{h : (h, \bar{x}) \in C'\}$, we have

$$\mathbb{P}_{\mathcal{Z}_t(\theta)}(C) = \int \pi_\theta(c_{\bar{x}}^{-1}(c_{\bar{x}}(C'_{\bar{x}}))) \mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) = \int \pi_\theta(C'_{\bar{x}}) \mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) = \mathbb{P}_{(h_{t\theta}^*, \mathbb{X})}(C').$$

Likewise, this reasoning holds for θ'. Then

$$
\begin{aligned}
\|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| &= \|\mathbb{P}_{(h_{t\theta}^*, \mathbb{X})} - \mathbb{P}_{(h_{t\theta'}^*, \mathbb{X})}\| \\
&= \sup_{C' \in \mathcal{B} \otimes \mathcal{B}_{\mathcal{X}}^\infty} \left| \int (\pi_\theta(C'_{\bar{x}}) - \pi_{\theta'}(C'_{\bar{x}})) \mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) \right| \\
&\leq \int \sup_{A \in \mathcal{B}} |\pi_\theta(A) - \pi_{\theta'}(A)| \mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) = \|\pi_\theta - \pi_{\theta'}\|.
\end{aligned}
$$

Since $h_{t\theta}^*$ and $\mathbb{X}$ are independent, for $A \in \mathcal{B}$, $\pi_\theta(A) = \mathbb{P}_{h_{t\theta}^*}(A) = \mathbb{P}_{h_{t\theta}^*}(A)\mathbb{P}_{\mathbb{X}}(\mathcal{X}^\infty) = \mathbb{P}_{(h_{t\theta}^*, \mathbb{X})}(A \times \mathcal{X}^\infty)$. Analogous reasoning holds for $h_{t\theta'}^*$. Thus, we have

$$
\begin{aligned}
\|\pi_\theta - \pi_{\theta'}\| &= \|\mathbb{P}_{(h_{t\theta}^*, \mathbb{X})}(\cdot \times \mathcal{X}^\infty) - \mathbb{P}_{(h_{t\theta'}^*, \mathbb{X})}(\cdot \times \mathcal{X}^\infty)\| \\
&\leq \|\mathbb{P}_{(h_{t\theta}^*, \mathbb{X})} - \mathbb{P}_{(h_{t\theta'}^*, \mathbb{X})}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\|.
\end{aligned}
$$

Combining the above, we have $\|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| = \|\pi_\theta - \pi_{\theta'}\|$. $\qquad\square$

Lemma 7.3. *There exists a sequence $r_k = o(1)$ such that $\forall t, k \in \mathbb{N}$, $\forall \theta, \theta' \in \Theta$,*

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq \|\pi_\theta - \pi_{\theta'}\| \leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + r_k.$$

Proof. The left inequality follows from Lemma 7.2 and the basic definition of $\|\cdot\|$, since $\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(\cdot) = \mathbb{P}_{\mathcal{Z}_t(\theta)}(\cdot \times (\mathcal{X} \times \{-1, +1\})^\infty)$, so that

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| = \|\pi_\theta - \pi_{\theta'}\|.$$

The remainder of this proof focuses on the right inequality. Fix $\theta, \theta' \in \Theta$, let $\gamma > 0$, and let $B \subseteq (\mathcal{X} \times \{-1, +1\})^\infty$ be a measurable set such that

$$\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| < \mathbb{P}_{\mathcal{Z}_t(\theta)}(B) - \mathbb{P}_{\mathcal{Z}_t(\theta')}(B) + \gamma.$$

Let $\mathcal{A}$ be the collection of all measurable subsets of $(\mathcal{X} \times \{-1, +1\})^\infty$ representable in the form $A' \times (\mathcal{X} \times \{-1, +1\})^\infty$, for some measurable $A' \subseteq (\mathcal{X} \times \{-1, +1\})^k$ and some $k \in \mathbb{N}$. In particular, since $\mathcal{A}$ is an algebra that generates the product σ-algebra, Carathéodory's extension theorem [Schervish, 1995] implies that there exist disjoint sets $\{A_i\}_{i \in \mathbb{N}}$ in $\mathcal{A}$ such that $B \subseteq \bigcup_{i \in \mathbb{N}} A_i$ and

$$\mathbb{P}_{\mathcal{Z}_t(\theta)}(B) - \mathbb{P}_{\mathcal{Z}_t(\theta')}(B) < \sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i) + \gamma.$$

Additionally, as these sums are bounded, there must exist $n \in \mathbb{N}$ such that

$$\sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) < \gamma + \sum_{i=1}^{n} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i),$$

so that

$$\sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i) < \gamma + \sum_{i=1}^{n} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i=1}^{n} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i)$$

$$= \gamma + \mathbb{P}_{\mathcal{Z}_t(\theta)}\left(\bigcup_{i=1}^{n} A_i\right) - \mathbb{P}_{\mathcal{Z}_t(\theta')}\left(\bigcup_{i=1}^{n} A_i\right).$$

121

As $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$, there exists $k' \in \mathbb{N}$ and measurable $A' \subseteq (\mathcal{X} \times \{-1, +1\})^{k'}$ such that $\bigcup_{i=1}^{n} A_i = A' \times (\mathcal{X} \times \{-1, +1\})^{\infty}$, and therefore

$$\mathbb{P}_{\mathcal{Z}_t(\theta)}\left(\bigcup_{i=1}^{n} A_i\right) - \mathbb{P}_{\mathcal{Z}_t(\theta')}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{P}_{\mathcal{Z}_{tk'}(\theta)}(A') - \mathbb{P}_{\mathcal{Z}_{tk'}(\theta')}(A')$$

$$\leq \|\mathbb{P}_{\mathcal{Z}_{tk'}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk'}(\theta')}\| \leq \lim_{k\to\infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.$$

In summary, we have $\|\pi_\theta - \pi_{\theta'}\| \leq \lim_{k\to\infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + 3\gamma$. Since this is true for an arbitrary $\gamma > 0$, taking the limit as $\gamma \to 0$ implies

$$\|\pi_\theta - \pi_{\theta'}\| \leq \lim_{k\to\infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.$$

In particular, this implies there exists a sequence $r_k(\theta, \theta') = o(1)$ such that

$$\forall k \in \mathbb{N}, \|\pi_\theta - \pi_{\theta'}\| \leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + r_k(\theta, \theta').$$

This would suffice to establish the upper bound if we were allowing r_k to depend on the particular θ and θ'. However, to guarantee the same rates of convergence for all pairs of parameters requires an additional argument. Specifically, let $\gamma > 0$ and let Θ_γ denote a minimal subset of Θ such that, $\forall \theta \in \Theta, \exists \theta_\gamma \in \Theta_\gamma$ s.t. $\|\pi_\theta - \pi_{\theta_\gamma}\| < \gamma$: that is, a minimal γ-cover. Since $|\Theta_\gamma| < \infty$ by assumption, defining $r_k(\gamma) = \max_{\theta, \theta' \in \Theta_\gamma} r_k(\theta, \theta')$, we have $r_k(\gamma) = o(1)$. Furthermore, for any $\theta, \theta' \in \Theta$, letting $\theta_\gamma = \mathrm{argmin}_{\theta'' \in \Theta_\gamma} \|\pi_\theta - \pi_{\theta''}\|$ and $\theta'_\gamma = \mathrm{argmin}_{\theta'' \in \Theta_\gamma} \|\pi_{\theta'} - \pi_{\theta''}\|$, we have (by triangle inequalities)

$$\|\pi_\theta - \pi_{\theta'}\| \leq \|\pi_\theta - \pi_{\theta_\gamma}\| + \|\pi_{\theta_\gamma} - \pi_{\theta'_\gamma}\| + \|\pi_{\theta'_\gamma} - \pi_{\theta'}\|$$

$$< 2\gamma + r_k(\gamma) + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\|.$$

By triangle inequalities and the left inequality from the lemma statement (established above), we

also have

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\|$$

$$\leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta')} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\|$$

$$\leq \|\pi_{\theta_\gamma} - \pi_\theta\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + \|\pi_{\theta'} - \pi_{\theta'_\gamma}\|$$

$$< 2\gamma + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.$$

Defining $r_k = \inf_{\gamma>0}(4\gamma + r_k(\gamma))$, we have the right inequality of the lemma statement, and since $r_k(\gamma) = o(1)$ for each $\gamma > 0$, we have $r_k = o(1)$. $\qquad\square$

Lemma 7.4. $\forall t, k \in \mathbb{N}, \forall \theta, \theta' \in \Theta,$

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq 4 \cdot 2^{2k+d} k^d \sqrt{\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|}.$$

Proof. Fix any $t \in \mathbb{N}$, and let $\mathbb{X} = \{X_{t1}, X_{t2}, \ldots\}$ and $\mathbb{Y}(\theta) = \{Y_{t1}(\theta), Y_{t2}(\theta), \ldots\}$, and for $k \in \mathbb{N}$ let $\mathbb{X}_k = \{X_{t1}, \ldots, X_{tk}\}$ and $\mathbb{Y}_k(\theta) = \{Y_{t1}(\theta), \ldots, Y_{tk}(\theta)\}$.

If $k \leq d$, then $\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(\cdot) = \mathbb{P}_{\mathcal{Z}_{td}(\theta)}(\cdot \times (\mathcal{X} \times \{-1, +1\})^{d-k})$, so that

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq \|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|,$$

and therefore the result trivially holds.

Now suppose $k > d$. For a sequence $\bar{z}$ and $I \subseteq \mathbb{N}$, we will use the notation $\bar{z}_I = \{\bar{z}_i : i \in I\}$. Note that, for any $k > d$ and $\bar{x}^k \in \mathcal{X}^k$, there is a sequence $\bar{y}(\bar{x}^k) \in \{-1, +1\}^k$ such that no $h \in \mathbb{C}$ has $h(\bar{x}^k) = \bar{y}(\bar{x}^k)$ (i.e., $\forall h \in \mathbb{C}, \exists i \leq k$ s.t. $h(\bar{x}_i^k) \neq \bar{y}_i(\bar{x}^k)$). Now suppose $k > d$ and take as an inductive hypothesis that there is a measurable set $A^* \subseteq \mathcal{X}^\infty$ of probability one with the property that $\forall \bar{x} \in A^*$, for every finite $I \subset \mathbb{N}$ with $|I| > d$, for every $\bar{y} \in \{-1, +1\}^\infty$ with $\|\bar{y}_I - \bar{y}(\bar{x}_I)\|_1 / 2 \leq k - 1$,

$$\left| \mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) \right|$$

$$\leq 2^{k-1} \cdot \max_{\tilde{y}^d \in \{-1,+1\}^d, D \in I^d} \left| \mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D) \right|.$$

This clearly holds for $\|\bar{y}_I - \bar{y}(\bar{x}_I)\|_1/2 = 0$, since $\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) = 0$ in this case, so this will serve as our base case in the inductive proof. Next we inductively extend this to the value $k > 0$. Specifically, let A^*_{k-1} be the A^* guaranteed to exist by the inductive hypothesis, and fix any $\bar{x} \in A^*$, $\bar{y} \in \{-1, +1\}^\infty$, and finite $I \subset \mathbb{N}$ with $|I| > d$ and $\|\bar{y}_I - \bar{y}(\bar{x}_I)\|_1/2 = k$. Let $i \in I$ be such that $\bar{y}_i \neq \bar{y}_i(\bar{x}_I)$, and let $\bar{y}' \in \{-1, +1\}$ have $\bar{y}'_j = \bar{y}_j$ for every $j \neq i$, and $\bar{y}'_i = -\bar{y}_i$. Then

$$\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) = \mathbb{P}_{\mathbb{Y}_{I\setminus\{i\}}(\theta)|\mathbb{X}_{I\setminus\{i\}}}(\bar{y}_{I\setminus\{i\}}|\bar{x}_{I\setminus\{i\}}) - \mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I), \tag{7.1}$$

and similarly for θ'. By the inductive hypothesis, this means

$$\left|\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I)\right|$$

$$\leq \left|\mathbb{P}_{\mathbb{Y}_{I\setminus\{i\}}(\theta)|\mathbb{X}_{I\setminus\{i\}}}(\bar{y}_{I\setminus\{i\}}|\bar{x}_{I\setminus\{i\}}) - \mathbb{P}_{\mathbb{Y}_{I\setminus\{i\}}(\theta')|\mathbb{X}_{I\setminus\{i\}}}(\bar{y}_{I\setminus\{i\}}|\bar{x}_{I\setminus\{i\}})\right|$$

$$+ \left|\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I)\right|$$

$$\leq 2^k \cdot \max_{\tilde{y}^d \in \{-1,+1\}^d, D \in I^d} \left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\bar{x}_D)\right|.$$

Therefore, by the principle of induction, this inequality holds for all $k > d$, for every $\bar{x} \in A^*$, $\bar{y} \in \{-1, +1\}^\infty$, and finite $I \subset \mathbb{N}$, where A^* has $\mathcal{D}^\infty$-probability one.

In particular, we have that for $\theta, \theta' \in \Theta$,

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|$$

$$\leq 2^k \mathbb{E}\left[\max_{\bar{y}^k \in \{-1,+1\}^k} \left|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k}(\bar{y}^k|\mathbb{X}_k) - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}(\bar{y}^k|\mathbb{X}_k)\right|\right]$$

$$\leq 2^{2k} \mathbb{E}\left[\max_{\tilde{y}^d \in \{-1,+1\}^d, D \in \{1,\ldots,k\}^d} \left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D)\right|\right]$$

$$\leq 2^{2k} \sum_{\tilde{y}^d \in \{-1,+1\}^d} \sum_{D \in \{1,\ldots,k\}^d} \mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_D)\right|\right].$$

Exchangeability implies this is at most

$$2^{2k} \sum_{\tilde{y}^d \in \{-1,+1\}^d} \sum_{D \in \{1,\ldots,k\}^d} \mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d)\right|\right]$$

$$\leq 2^{2k+d} k^d \max_{\tilde{y}^d \in \{-1,+1\}^d} \mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d)\right|\right].$$

124

To complete the proof, we need only bound this value by an appropriate function of $\|\mathbb{P}_{\mathcal{Z}_{td(\theta)}} - \mathbb{P}_{\mathcal{Z}_{td(\theta')}}\|$. Toward this end, suppose

$$\mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_{d(\theta)}|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_{d(\theta')}|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d)\right|\right] \geq \varepsilon,$$

for some $\tilde{y}^d$. Then either

$$\mathbb{P}\left(\mathbb{P}_{\mathbb{Y}_{d(\theta)}|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_{d(\theta')}|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) \geq \varepsilon/4\right) \geq \varepsilon/4,$$

or

$$\mathbb{P}\left(\mathbb{P}_{\mathbb{Y}_{d(\theta')}|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_{d(\theta)}|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) \geq \varepsilon/4\right) \geq \varepsilon/4.$$

For which ever is the case, let A_ε denote the corresponding measurable subset of $\mathcal{X}^d$, of probability at least $\varepsilon/4$. Then

$$\|\mathbb{P}_{\mathcal{Z}_{td(\theta)}} - \mathbb{P}_{\mathcal{Z}_{td(\theta')}}\| \geq \left|\mathbb{P}_{\mathcal{Z}_{td(\theta)}}(A_\varepsilon \times \{\tilde{y}^d\}) - \mathbb{P}_{\mathcal{Z}_{td(\theta')}}(A_\varepsilon \times \{\tilde{y}^d\})\right|$$

$$\geq (\varepsilon/4)\mathbb{P}_{\mathbb{X}_d}(A_\varepsilon) \geq \varepsilon^2/16.$$

Therefore,

$$\mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_{d(\theta)}|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_{d(\theta')}|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d)\right|\right] \leq 4\sqrt{\|\mathbb{P}_{\mathcal{Z}_{td(\theta)}} - \mathbb{P}_{\mathcal{Z}_{td(\theta')}}\|},$$

which means

$$2^{2k+d}k^d \max_{\tilde{y}^d \in \{-1,+1\}^d} \mathbb{E}\left[\left|\mathbb{P}_{\mathbb{Y}_{d(\theta)}|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d) - \mathbb{P}_{\mathbb{Y}_{d(\theta')}|\mathbb{X}_d}(\tilde{y}^d|\mathbb{X}_d)\right|\right]$$

$$\leq 4 \cdot 2^{2k+d}k^d \sqrt{\|\mathbb{P}_{\mathcal{Z}_{td(\theta)}} - \mathbb{P}_{\mathcal{Z}_{td(\theta')}}\|}.$$

$\square$

The following lemma is a standard result on the existence of converging density estimators for totally bounded families of distributions. For our purposes, the details of the estimator achieving this guarantee are not particularly important, as we will apply the result as stated. For completeness, we describe a particular estimator that does achieve the guarantee after the lemma.

Lemma 7.5. *[Devroye and Lugosi, 2001, Yatracos, 1985] Let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ be a totally bounded family of probability measures on a measurable space $(\Omega, \mathcal{F})$, and let $\{W_t(\theta)\}_{t \in \mathbb{N}, \theta \in \Theta}$ be Ω-valued random variables such that $\{W_t(\theta)\}_{t \in \mathbb{N}}$ are i.i.d. p_θ for each $\theta \in \Theta$. Then there exists an estimator $\hat{\theta}_{T\theta_\star} = \hat{\theta}_T(W_1(\theta_\star), \dots, W_T(\theta_\star))$ and functions $R_\mathcal{P} : \mathbb{N}_0 \times (0,1] \to [0, \infty)$ and $\delta_\mathcal{P} : \mathbb{N}_0 \times (0,1] \to [0,1]$ such that $\forall \alpha > 0$, $\lim_{T\to\infty} R_\mathcal{P}(T, \alpha) = \lim_{T\to\infty} \delta_\mathcal{P}(T, \alpha) = 0$, and $\forall \theta_\star \in \Theta$ and $T \in \mathbb{N}_0$,*

$$\mathbb{P}\left(\|p_{\hat{\theta}_{T\theta_\star}} - p_{\theta_\star}\| > R_\mathcal{P}(T, \alpha) \right) \leq \delta_\mathcal{P}(T, \alpha) \leq \alpha.$$

In many contexts (though certainly not all), even a simple maximum likelihood estimator suffices to supply this guarantee. However, to derive results under the more general conditions we consider here, we require a more involved method: specifically, the minimum distance skeleton estimate explored by [Devroye and Lugosi, 2001, Yatracos, 1985], specified as follows. Let $\Theta_\varepsilon \subseteq \Theta$ be a minimal-cardinality ε-cover of Θ: that is, a minimal-cardinality subset of Θ such that $\forall \theta \in \Theta$, $\exists \theta_\varepsilon \in \Theta_\varepsilon$ with $\|p_{\theta_\varepsilon} - p_\theta\| < \varepsilon$. For each $\theta, \theta' \in \Theta_\varepsilon$, let $A_{\theta,\theta'}$ be a set in $\mathcal{F}$ maximizing $p_\theta(A_{\theta,\theta'}) - p_{\theta'}(A_{\theta,\theta'})$, and let $\mathcal{A}_\varepsilon = \{A_{\theta,\theta'} : \theta, \theta' \in \Theta_\varepsilon\}$, known as a *Yatracos class*. Finally, for $A \in \mathcal{F}$, let $\hat{p}_T(A) = T^{-1} \sum_{t=1}^{T} \mathbb{1}_A(W_t(\theta_\star))$. The minimum distance skeleton estimate is $\hat{\theta}_{T\theta_\star} = \operatorname{argmin}_{\theta \in \Theta_\varepsilon} \sup_{A \in \mathcal{A}_\varepsilon} |p_\theta(A) - \hat{p}_T(A)|$. The reader is referred to [Devroye and Lugosi, 2001, Yatracos, 1985] for a proof that this method satisfies the guarantee of Lemma 7.5. In particular, if ε_T is a sequence decreasing to 0 at a rate such that $T^{-1} \log(|\Theta_{\varepsilon_T}|) \to 0$, and δ_T is a sequence bounded by α and decreasing to 0 with $\delta_T = \omega(\varepsilon_T + \sqrt{T^{-1} \log(|\Theta_{\varepsilon_T}|)})$, then the result of [Devroye and Lugosi, 2001, Yatracos, 1985], combined with Markov's inequality, implies that to satisfy the condition of Lemma 7.5, it suffices to take $R_\mathcal{P}(T, \alpha) = \delta_T^{-1} \left(3\varepsilon_T + \sqrt{8T^{-1} \log(2|\Theta_{\varepsilon_T}|^2 \vee 8)} \right)$ and $\delta_\mathcal{P}(T, \alpha) = \delta_T$. For instance, $\varepsilon_T = 2 \inf \left\{ \varepsilon > 0 : \log(|\Theta_\varepsilon|) \leq \sqrt{T} \right\}$ and $\delta_T = \alpha \wedge (\sqrt{\varepsilon_T} + T^{-1/8})$ suffice.

We are now ready for the proof of Theorem 7.1

Theorem 7.1. For $\varepsilon > 0$, let $\Theta_\varepsilon \subseteq \Theta$ be a finite subset such that $\forall \theta \in \Theta$, $\exists \theta_\varepsilon \in \Theta_\varepsilon$ with $\|\pi_{\theta_\varepsilon} - \pi_\theta\| < \varepsilon$; this exists by the assumption that $\{\pi_\theta : \theta \in \Theta\}$ is totally bounded. Then

126

Lemma 7.3 implies that $\forall \theta \in \Theta$, $\exists \theta_\varepsilon \in \Theta_\varepsilon$ with $\|\mathbb{P}_{\mathcal{Z}_{td}(\theta_\varepsilon)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta)}\| \leq \|\pi_{\theta_\varepsilon} - \pi_\theta\| < \varepsilon$, so that $\{\mathbb{P}_{\mathcal{Z}_{td}(\theta_\varepsilon)} : \theta_\varepsilon \in \Theta_\varepsilon\}$ is a finite ε-cover of $\{\mathbb{P}_{\mathcal{Z}_{td}(\theta)} : \theta \in \Theta\}$. Therefore, $\{\mathbb{P}_{\mathcal{Z}_{td}(\theta)} : \theta \in \Theta\}$ is totally bounded. Lemma 7.5 then implies that there exists an estimator $\hat{\theta}_{T\theta_\star} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star))$ and functions $R_d : \mathbb{N}_0 \times (0,1] \to [0,\infty)$ and $\delta_d : \mathbb{N}_0 \times (0,1] \to [0,1]$ such that $\forall \alpha > 0$, $\lim_{T \to \infty} R_d(T, \alpha) = \lim_{T \to \infty} \delta_d(T, \alpha) = 0$, and $\forall \theta_\star \in \Theta$ and $T \in \mathbb{N}_0$,

$$\mathbb{P}\left(\|\mathbb{P}_{\mathcal{Z}_{(T+1)d}(\hat{\theta}_{T\theta_\star})|\hat{\theta}_{T\theta_\star}} - \mathbb{P}_{\mathcal{Z}_{(T+1)d}(\theta_\star)}\| > R_d(T, \alpha)\right) \leq \delta_d(T, \alpha) \leq \alpha. \tag{7.2}$$

Defining

$$R(T, \alpha) = \min_{k \in \mathbb{N}}\left(r_k + 4 \cdot 2^{2k+d} k^d \sqrt{R_d(T, \alpha)}\right),$$

and $\delta(T, \alpha) = \delta_d(T, \alpha)$, and combining (7.2) with Lemmas 7.4 and 7.3, we have

$$\mathbb{P}\left(\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\| > R(T, \alpha)\right) \leq \delta(T, \alpha) \leq \alpha.$$

Finally, note that $\lim_{k \to \infty} r_k = 0$ and $\lim_{T \to \infty} R_d(T, \alpha) = 0$ imply that $\lim_{T \to \infty} R(T, \alpha) = 0$. $\square$

7.3.1 Identifiability from d Points

Inspection of the above proof reveals that the assumption that the family of priors is totally bounded is required only to establish the estimability and bounded minimax rate guarantees. In particular, the implied identifiability condition is, in fact, *always* satisfied, as stated formally in the following corollary.

Corollary 7.6. *For any priors π_1, π_2 on $\mathbb{C}$, if $h_i^* \sim \pi_i$, $X_1, \ldots, X_d$ are i.i.d. $\mathcal{D}$ independent from h_i^*, and $Z_d(i) = \{(X_1, h_i^*(X_1)), \ldots, (X_d, h_i^*(X_d))\}$ for $i \in \{1, 2\}$, then $\mathbb{P}_{Z_d(1)} = \mathbb{P}_{Z_d(2)} \implies \pi_1 = \pi_2$.*

Proof. The described scenario is a special case of our general setting, with $\Theta = \{1, 2\}$, in which case $\mathbb{P}_{Z_d(i)} = \mathbb{P}_{\mathcal{Z}_{1d}(i)}$. Thus, if $\mathbb{P}_{Z_d(1)} = \mathbb{P}_{Z_d(2)}$, then Lemma 7.4 and Lemma 7.3 combine to imply that $\|\pi_1 - \pi_2\| \leq \inf_{k \in \mathbb{N}} r_k = 0$. $\square$

Since Corollary 7.6 is interesting in itself, it is worth noting that there is a simple direct proof of this result. Specifically, by an inductive argument based on the observation (7.1) from the proof of Lemma 7.4, we quickly find that for any $k \in \mathbb{N}$, $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$ is identifiable from $\mathbb{P}_{\mathcal{Z}_{td}(\theta_\star)}$. Then we merely recall that $\mathbb{P}_{\mathcal{Z}_t(\theta_\star)}$ is always identifiable from $\{\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)} : k \in \mathbb{N}\}$ [Kallenberg, 2002], and the argument from the proof of Lemma 7.2 shows $\pi_{\theta_\star}$ is identifiable from $\mathbb{P}_{\mathcal{Z}_t(\theta_\star)}$.

It is natural to wonder whether identifiability of $\pi_{\theta_\star}$ from $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$ remains true for some smaller number of points $k < d$, so that we might hope to create an estimator for $\pi_{\theta_\star}$ based on an estimator for $\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\star)}$. However, one can show that d is actually the *minimum* possible value for which this remains true for all $\mathcal{D}$ and all families of priors. Formally, we have the following result, holding for every VC class $\mathbb{C}$.

Theorem 7.7. *There exists a data distribution $\mathcal{D}$ and priors π_1, π_2 on $\mathbb{C}$ such that, for any positive integer $k < d$, if $h_i^* \sim \pi_i$, $X_1, \ldots, X_k$ are i.i.d. $\mathcal{D}$ independent from h_i^*, and $Z_k(i) = \{(X_1, h_i^*(X_1)), \ldots, (X_k, h_i^*(X_k))\}$ for $i \in \{1, 2\}$, then $\mathbb{P}_{Z_k(1)} = \mathbb{P}_{Z_k(2)}$ but $\pi_1 \neq \pi_2$.*

Proof. Note that it suffices to show this is the case for $k = d - 1$, since any smaller k is a marginal of this case. Consider a shatterable set of points $S_d = \{x_1, x_2, \ldots, x_d\} \subseteq \mathcal{X}$, and let $\mathcal{D}$ be uniform on S_d. Let $\mathbb{C}[S_d]$ be any 2^d classifiers in $\mathbb{C}$ that shatter S_d. Let π_1 be the uniform distribution on $\mathbb{C}[S]$. Now let $S_{d-1} = \{x_1, \ldots, x_{d-1}\}$ and $\mathbb{C}[S_{d-1}] \subseteq \mathbb{C}[S_d]$ shatter S_{d-1} with the property that $\forall h \in \mathbb{C}[S_{d-1}]$, $h(x_d) = \prod_{j=1}^{d-1} h(x_j)$. Let π_2 be uniform on $\mathbb{C}[S_{d-1}]$. Now for any $k < d$ and distinct indices $t_1, \ldots, t_k \in \{1, \ldots, d\}$, $\{h_i^*(x_{t_1}), \ldots, h_i^*(x_{t_k})\}$ is distributed uniformly in $\{-1, +1\}^k$ for both $i \in \{1, 2\}$. This implies $\mathbb{P}_{Z_{d-1}(1)|X_1,\ldots,X_{d-1}} = \mathbb{P}_{Z_{d-1}(2)|X_1,\ldots,X_{d-1}}$, which implies $\mathbb{P}_{Z_{d-1}(1)} = \mathbb{P}_{Z_{d-1}(2)}$. However, π_1 is clearly different from π_2, since even the sizes of the supports are different. $\square$

7.4 Transfer Learning

In this section, we look at an application of the techniques from the previous section to transfer learning. Like the previous section, the results in this section are general, in that they are applicable to a variety of learning protocols, including passive supervised learning, passive semi-supervised learning, active learning, and learning with certain general types of data-dependent interaction (see [Hanneke, 2009]). For simplicity, we restrict our discussion to the active learning formulation; the analogous results for these other learning protocols follow by similar reasoning.

The result of the previous section implies that an estimator for $\theta_\star$ based on d-dimensional joint distributions is consistent with a bounded rate of convergence R. Therefore, for certain prior-dependent learning algorithms, their behavior should be similar under $\pi_{\hat{\theta}_{T\theta_\star}}$ to their behavior under $\pi_{\theta_\star}$.

To make this concrete, we formalize this in the active learning protocol as follows. A *prior-dependent* active learning algorithm $\mathcal{A}$ takes as inputs $\varepsilon > 0$, $\mathcal{D}$, and a distribution π on $\mathbb{C}$. It initially has access to $X_1, X_2, \ldots$ i.i.d. $\mathcal{D}$; it then selects an index i_1 to request the label for, receives $Y_{i_1} = h^*(X_{i_1})$, then selects another index i_2, etc., until it eventually terminates and returns a classifier. Denote by $\mathcal{Z} = \{(X_1, h^*(X_1)), (X_2, h^*(X_2)), \ldots\}$. To be *correct*, $\mathcal{A}$ must guarantee that for $h^* \sim \pi$, $\forall \varepsilon > 0$, $\mathbb{E}\left[\rho(\mathcal{A}(\varepsilon, \mathcal{D}, \pi), h^*)\right] \leq \varepsilon$. We define the random variable $N(\mathcal{A}, f, \varepsilon, \mathcal{D}, \pi)$ as the number of label requests $\mathcal{A}$ makes before terminating, when given ε, $\mathcal{D}$, and π as inputs, and when $h^* = f$ is the value of the target function; we make the particular data sequence $\mathcal{Z}$ the algorithm is run with implicit in this notation. We will be interested in the *expected sample complexity* $SC(\mathcal{A}, \varepsilon, \mathcal{D}, \pi) = \mathbb{E}\left[N(\mathcal{A}, h^*, \varepsilon, \mathcal{D}, \pi)\right]$.

We propose the following algorithm $\mathcal{A}_\tau$ for transfer learning, defined in terms of a given correct prior-dependent active learning algorithm $\mathcal{A}_a$. We discuss interesting specifications for $\mathcal{A}_a$ in the next section, but for now the only assumption we require is that for any $\varepsilon > 0$ and $\mathcal{D}$, there is a value $s_\varepsilon < \infty$ such that for every π and $f \in \mathbb{C}$, $N(\mathcal{A}_a, f, \varepsilon, \mathcal{D}, \pi) \leq s_\varepsilon$; this is a very mild requirement, and any active learning algorithm can be converted into one that

satisfies this without significantly increasing its sample complexities for the priors it is already good for [Balcan, Hanneke, and Vaughan, 2010]. We additionally denote by $m_\varepsilon = \frac{16d}{\varepsilon} \ln\left(\frac{24}{\varepsilon}\right)$, and $\mathrm{B}(\theta, \gamma) = \{\theta' \in \Theta : \|\pi_\theta - \pi_{\theta'}\| \leq \gamma\}$.

Algorithm 1 $\mathcal{A}_\tau(T, \varepsilon)$: an algorithm for transfer learning, specified in terms of a generic subroutine $\mathcal{A}_a$.

---

for $t = 1, 2, \ldots, T$ **do**

    Request labels $Y_{t1}(\theta_\star), \ldots, Y_{td}(\theta_\star)$

    if $R(t - 1, \varepsilon/2) > \varepsilon/8$ **then**

        Request labels $Y_{t(d+1)}(\theta_\star), \ldots, Y_{tm_\varepsilon}(\theta_\star)$

        Take $\hat{h}_t$ as any $h \in \mathbb{C}$ s.t. $\forall i \leq m_\varepsilon$, $h(X_{ti}) = Y_{ti}(\theta_\star)$

    else

        Let $\breve{\theta}_{t\theta_\star} \in \mathrm{B}\left(\hat{\theta}_{(t-1)\theta_\star}, R(t - 1, \varepsilon/2)\right)$ be such that

$$SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\breve{\theta}_{t\theta_\star}}) \leq \min_{\theta \in \mathrm{B}\left(\hat{\theta}_{(t-1)\theta_\star}, R(t-1, \varepsilon/2)\right)} SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_\theta) + 1/t$$

        Run $\mathcal{A}_a(\varepsilon/4, \mathcal{D}, \pi_{\breve{\theta}_{t\theta_\star}})$ with data sequence $\mathcal{Z}_t(\theta_\star)$ and let $\hat{h}_t$ be the classifier it returns

    end if

end for

Recall that $\hat{\theta}_{(t-1)\theta_\star}$, which is defined by Theorem 7.1, is a function of the labels requested on previous rounds of the algorithm; $R(t - 1, \varepsilon/2)$ is also defined by Theorem 7.1, and has no dependence on the data (or on $\theta_\star$). The other quantities referred to in Algorithm 1 are defined just prior to Algorithm 1. We suppose the algorithm has access to the value $SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_\theta)$ for every $\theta \in \Theta$. This can sometimes be calculated analytically as a function of θ, or else can typically be approximated via Monte Carlo simulations. In fact, the result below holds even if SC is merely an accessible *upper bound* on the expected sample complexity.

Theorem 7.8. *The algorithm $\mathcal{A}_\tau$ is correct. Furthermore, if $S_T(\varepsilon)$ is the total number of label requests made by $\mathcal{A}_\tau(T, \varepsilon)$, then* $\limsup_{T \to \infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq SC(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_\star}) + d.$

The implication of Theorem 7.8 is that, via transfer learning, it is possible to achieve al-

most the *same* long-run average sample complexity as would be achievable if the target's prior distribution were *known* to the learner. We will see in the next section that this is sometimes significantly better than the single-task sample complexity. As mentioned, results of this type for transfer learning have previously appeared when $\mathcal{A}_a$ is a passive learning method [Baxter, 1997]; however, to our knowledge, this is the first such result where the asymptotics concern only the number of learning tasks, not the number of samples per task; this is also the first result we know of that is immediately applicable to more sophisticated learning protocols such as active learning.

The algorithm $\mathcal{A}_\tau$ is stated in a simple way here, but Theorem 7.8 can be improved with some obvious modifications to $\mathcal{A}_\tau$. The extra "$+d$" in Theorem 7.8 is not actually necessary, since we could stop updating the estimator $\check{\theta}_{t\theta_\star}$ (and the corresponding R value) after some $o(T)$ number of rounds (e.g., $\sqrt{T}$), in which case we would not need to request $Y_{t1}(\theta_\star), \ldots, Y_{td}(\theta_\star)$ for t larger than this, and the extra $d \cdot o(T)$ number of labeled examples vanishes in the average as $T \to \infty$. Additionally, the $\varepsilon/4$ term can easily be improved to any value arbitrarily close to ε (even $(1 - o(1))\varepsilon$) by running $\mathcal{A}_a$ with argument $\varepsilon - 2R(t - 1, \varepsilon/2) - \delta(t - 1, \varepsilon/2)$ instead of $\varepsilon/4$, and using this value in the SC calculations in the definition of $\check{\theta}_{t\theta_\star}$ as well. In fact, for many algorithms $\mathcal{A}_a$ (e.g., with $SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi_{\theta_\star})$ continuous in ε), combining the above two tricks yields $\limsup\limits_{T \to \infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq SC(\mathcal{A}_a, \varepsilon, \mathcal{D}, \pi_{\theta_\star})$.

Returning to our motivational remarks from Subsection 7.2.1, we can ask how many *extra* labeled examples are required from each learning problem to gain the benefits of transfer learning. This question essentially concerns the initial step of requesting the labels $Y_{t1}(\theta_\star), \ldots, Y_{td}(\theta_\star)$. Clearly this indicates that from each learning problem, we need at most d extra labeled examples to gain the benefits of transfer. Whether these d label requests are indeed *extra* depends on the particular learning algorithm $\mathcal{A}_a$; that is, in some cases (e.g., certain passive learning algorithms), $\mathcal{A}_a$ may itself use these initial d labels for learning, so that in these cases the benefits of transfer learning are essentially gained as a *by-product* of the learning processes, and essentially no additional labeling effort need be expended to gain these benefits. On the other hand, for some

active learning algorithms, we may expect that at least some of these initial d labels would not be requested by the algorithm, so that some extra labeling effort is expended to gain the benefits of transfer in these cases.

One drawback of our approach is that we require the data distribution $\mathcal{D}$ to remain fixed across tasks (this contrasts with [Baxter, 1997]). However, it should be possible to relax this requirement in the active learning setting in many cases. For instance, if $\mathcal{X} = \mathbb{R}^k$, then as long as we are guaranteed that the distribution $\mathcal{D}_t$ for each learning task has a strictly positive density function, it should be possible to use rejection sampling for each task to guarantee the d queried examples from each task have approximately the same distribution across tasks. This is all we require for our consistency results on $\hat{\theta}_{T\theta_\star}$ (i.e., it was not important that the d samples came from the true distribution $\mathcal{D}$, only that they came from a distribution under which ρ is a metric). We leave the details of such an adaptive method for future consideration.

7.4.1 Proof of Theorem 7.8

Recall that, to establish correctness, we must show that $\forall t \leq T$, $\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\right] \leq \varepsilon$, regardless of the value of $\theta_\star \in \Theta$. Fix any $\theta_\star \in \Theta$ and $t \leq T$. If $R(t-1, \varepsilon/2) > \varepsilon/8$, then classic results from passive learning indicate that $\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\right] \leq \varepsilon$ [Vapnik, 1982]. Otherwise, by Theorem 7.1, with probability at least $1 - \varepsilon/2$, we have $\|\pi_{\theta_\star} - \pi_{\hat{\theta}_{(t-1)\theta_\star}}\| \leq R(t-1, \varepsilon/2)$. On this event, if $R(t-1, \varepsilon/2) \leq \varepsilon/8$, then by a triangle inequality $\|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq 2R(t-1, \varepsilon/2) \leq \varepsilon/4$. Thus,

$$\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\right] \leq \mathbb{E}\left[\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\bigg|\check{\theta}_{t\theta_\star}\right]\mathbb{1}\left[\|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq \varepsilon/4\right]\right] + \varepsilon/2. \tag{7.3}$$

For $\theta \in \Theta$, let $\hat{h}_{t\theta}$ denote the classifier that would be returned by $\mathcal{A}_a(\varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}})$ when run with data sequence $\{(X_{t1}, h^*_{t\theta}(X_{t1})), (X_{t2}, h^*_{t\theta}(X_{t2})), \ldots\}$. Note that for any $\theta \in \Theta$, any measurable function $F : \mathbb{C} \to [0, 1]$ has

$$\mathbb{E}\left[F(h^*_{t\theta_\star})\right] \leq \mathbb{E}\left[F(h^*_{t\theta})\right] + \|\pi_\theta - \pi_{\theta_\star}\|. \tag{7.4}$$

In particular, supposing $\|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq \varepsilon/4$, we have

$$\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\Big|\check{\theta}_{t\theta_\star}\right] = \mathbb{E}\left[\rho\left(\hat{h}_{t\theta_\star}, h^*_{t\theta_\star}\right)\Big|\check{\theta}_{t\theta_\star}\right]$$

$$\leq \mathbb{E}\left[\rho\left(\hat{h}_{t\check{\theta}_{t\theta_\star}}, h^*_{t\check{\theta}_{t\theta_\star}}\right)\Big|\check{\theta}_{t\theta_\star}\right] + \|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq \varepsilon/4 + \varepsilon/4 = \varepsilon/2.$$

Combined with (7.3), this implies $\mathbb{E}\left[\rho\left(\hat{h}_t, h^*_{t\theta_\star}\right)\right] \leq \varepsilon$.

We establish the sample complexity claim as follows. First note that convergence of $R(t - 1, \varepsilon/2)$ implies that $\lim_{T\to\infty}\sum_{t=1}^T \mathbb{1}\left[R(t, \varepsilon/2) > \varepsilon/8\right]/T = 0$, and that the number of labels used for a value of t with $R(t-1, \varepsilon/2) > \varepsilon/8$ is bounded by a finite function m_ε of ε. Therefore,

$$\limsup_{T\to\infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T}$$

$$\leq d + \limsup_{T\to\infty} \sum_{t=1}^T \mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\theta_\star}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}})\right]\mathbb{1}\left[R(t-1, \varepsilon/2) \leq \varepsilon/8\right]/T$$

$$\leq d + \limsup_{T\to\infty} \sum_{t=1}^T \mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\theta_\star}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}})\right]/T. \tag{7.5}$$

By the definition of R, δ from Theorem 7.1, we have

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\theta_\star}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}})\mathbb{1}\left[\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| > R(t-1, \varepsilon/2)\right]\right]$$

$$\leq \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T s_{\varepsilon/4}\mathbb{P}\left(\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| > R(t-1, \varepsilon/2)\right)$$

$$\leq s_{\varepsilon/4}\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \delta(t-1, \varepsilon/2) = 0.$$

Combined with (7.5), this implies

$$\limsup_{T\to\infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq d+$$

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\theta_\star}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}})\mathbb{1}\left[\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| \leq R(t-1, \varepsilon/2)\right]\right].$$

For any $t \leq T$, on the event $\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| \leq R(t-1, \varepsilon/2)$, we have (by the property (7.4)

133

and a triangle inequality)

$$\mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\theta_\star}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}}) \Big| \check{\theta}_{t\theta_\star}\right]$$

$$\leq \mathbb{E}\left[N(\mathcal{A}_a, h^*_{t\check{\theta}_{t\theta_\star}}, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}}) \Big| \check{\theta}_{t\theta_\star}\right] + 2R(t-1, \varepsilon/2)$$

$$= SC\left(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\check{\theta}_{t\theta_\star}}\right) + 2R(t-1, \varepsilon/2)$$

$$\leq SC\left(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_\star}\right) + 1/t + 2R(t-1, \varepsilon/2),$$

where the last inequality follows by definition of $\check{\theta}_{t\theta_\star}$. Therefore,

$$\limsup_{T \to \infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T}$$

$$\leq d + \limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} SC\left(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_\star}\right) + 1/t + 2R(t-1, \varepsilon/2)$$

$$= d + SC\left(\mathcal{A}_a, \varepsilon/4, \mathcal{D}, \pi_{\theta_\star}\right).$$

$\square$

7.5 Conclusions

We have shown that when learning a sequence of i.i.d. target concepts from a known VC class, with an unknown distribution from a known totally bounded family, transfer learning can lead to amortized average sample complexity close to that achievable by an algorithm with direct knowledge of the the targets' distribution.

Chapter 8

Prior Estimation

Abstract

[1]We study the optimal rates of convergence for estimating a prior distribution over a VC class from a sequence of independent data sets respectively labeled by independent target functions sampled from the prior. We specifically derive upper and lower bounds on the optimal rates under a smoothness condition on the correct prior, with the number of samples per data set equal the VC dimension. These results have implications for the improvements achievable via transfer learning.

8.1   Introduction

In the *transfer learning* setting, we are presented with a sequence of learning problems, each with some respective target concept we are tasked with learning. The key question in transfer learning is how to leverage our access to past learning problems in order to improve performance on learning problems we will be presented with in the future.

Among the several proposed models for transfer learning, one particularly appealing model

[1]Joint work with Jaime Carbonell and Steve Hanneke

supposes the learning problems are independent and identically distributed, with unknown distribution, and the advantage of transfer learning then comes from the ability to estimate this shared distribution based on the data from past learning problems [Baxter, 1997, Yang, Hanneke, and Carbonell, 2011]. For instance, when customizing a speech recognition system to a particular speaker's voice, we might expect the first few people would need to speak many words or phrases in order for the system to accurately identify the nuances. However, after performing this for many different people, if the software has access to those past training sessions when customizing itself to a new user, it should have identified important properties of the speech patterns, such as the common patterns within each of the major dialects or accents, and other such information about the *distribution* of speech patterns within the user population. It should then be able to leverage this information to reduce the number of words or phrases the next user needs to speak in order to train the system, for instance by first trying to identify the individual's dialect, then presenting phrases that differentiate common subpatterns within that dialect, and so forth.

In analyzing the benefits of transfer learning in such a setting, one important question to ask is how quickly we can estimate the distribution from which the learning problems are sampled. In recent work, Yang, Hanneke, and Carbonell [2011] have shown that under mild conditions on the family of possible distributions, if the target concepts reside in a known VC class, then it is possible to estimate this distribtion using only a bounded number of training samples per task: specifically, a number of samples equal the VC dimension. However, we left open the question of quantifying the *rate* of convergence. This rate of convergence can have a direct impact on how much benefit we gain from transfer learning when we are faced with only a finite sequence of learning problems. As such, it is certainly desirable to derive tight characterizations of this rate of convergence.

The present work continues that of Yang, Hanneke, and Carbonell [2011], bounding the rate of convergence for estimating this distribution, under a smoothness condition on the distribution.

We derive a generic upper bound, which holds regardless of the VC class the target concepts reside in. The proof of this result builds on our earlier work, but requires several interesting innovations to make the rate of convergence explicit, and to dramatically improve the upper bound implicit in the proofs of those earlier results. We further derive a nontrivial lower bound that holds for certain constructed scenarios, which illustrates a lower limit on how good of a general upper bound we might hope for in results expressed only in terms of the number of tasks, the smoothness conditions, and the VC dimension.

8.2   The Setting

Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a Borel space [Schervish, 1995] (where $\mathcal{X}$ is called the *instance space*), and let $\mathcal{D}$ be a distribution on $\mathcal{X}$ (called the *data distribution*). Let $\mathbb{C}$ be a VC class of measurable classifiers $h : \mathcal{X} \to \{-1, +1\}$ (called the *concept space*), and denote by d the VC dimension of $\mathbb{C}$ [Vapnik, 1982]. We suppose $\mathbb{C}$ is equipped with its Borel σ-algebra $\mathcal{B}$ induced by the pseudo-metric $\rho(h, g) = \mathcal{D}(\{x \in \mathcal{X} : h(x) \neq g(x)\})$. Though our results can be formulated for general $\mathcal{D}$ (with somewhat more complicated theorem statements), to simplify the statement of results we suppose ρ is actually a *metric*, which would follow from appropriate topological conditions on $\mathbb{C}$ relative to $\mathcal{D}$. For any two probability measures μ_1, μ_2 on a measurable space $(\Omega, \mathcal{F})$, define the total variation distance

$$\|\mu_1 - \mu_2\| = \sup_{A \in \mathcal{F}} \mu_1(A) - \mu_2(A).$$

Let $\Pi_{\Theta} = \{\pi_{\theta} : \theta \in \Theta\}$ be a family of probability measures on $\mathbb{C}$ (called *priors*), where Θ is an arbitrary index set (called the *parameter space*). We additionally suppose there exists a probability measure π_0 on $\mathbb{C}$ (called the *reference measure*) such that every π_{θ} is absolutely continuous with respect to π_0, and therefore has a density function f_{θ} given by the Radon-Nikodym derivative $\frac{d\pi_{\theta}}{d\pi_0}$ [Schervish, 1995].

We consider the following type of estimation problem. There is a collection of $\mathbb{C}$-valued ran-

dom variables $\{h_{t\theta}^* : t \in \mathbb{N}, \theta \in \Theta\}$, where for any fixed $\theta \in \Theta$ the $\{h_{t\theta}^*\}_{t=1}^\infty$ variables are i.i.d. with distribution π_θ. For each $\theta \in \Theta$, there is a sequence $\mathcal{Z}_t(\theta) = \{(X_{t1}, Y_{t1}(\theta)), (X_{t2}, Y_{t2}(\theta)), \ldots\}$, where $\{X_{ti}\}_{t,i\in\mathbb{N}}$ are i.i.d. $\mathcal{D}$, and for each $t, i \in \mathbb{N}$, $Y_{ti}(\theta) = h_{t\theta}^*(X_{ti})$. We additionally denote by $\mathcal{Z}_{tk} = \{(X_{t1}, Y_{t1}(\theta)), \ldots, (X_{tk}, Y_{tk}(\theta))\}$ the first k elements of $\mathcal{Z}_t(\theta)$, for any $k \in \mathbb{N}$, and similarly $\mathbb{X}_{tk} = \{X_{t1}, \ldots, X_{tk}\}$ and $\mathbb{Y}_{tk}(\theta) = \{Y_{t1}(\theta), \ldots, Y_{tk}(\theta)\}$. Following the terminology used in the transfer learning literature, we refer to the collection of variables associated with each t collectively as the tth *task*. We will be concerned with sequences of estimators $\hat{\theta}_{T\theta} = \hat{\theta}_T(\mathcal{Z}_{1k}(\theta), \ldots, \mathcal{Z}_{Tk}(\theta))$, for $T \in \mathbb{N}$, which are based on only a bounded number k of samples per task, among the first T tasks. Our main results specifically study the case of $k = d$. For any such estimator, we measure the *risk* as $\mathbb{E}\left[\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\|\right]$, and will be particularly interested in upper-bounding the worst-case risk $\sup_{\theta_\star \in \Theta} \mathbb{E}\left[\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\|\right]$ as a function of T, and lower-bounding the minimum possible value of this worst-case risk over all possible $\hat{\theta}_T$ estimators (called the *minimax risk*).

In previous work, Yang, Hanneke, and Carbonell [2011] we showed that, if Π_Θ is a totally bounded family, then even with only d number of samples per task, the minimax risk (as a function of the number of tasks T) converges to zero. In fact, we also proved this is not necessarily the case in general for any number of samples less than d. However, the actual rates of convergence were not explicitly derived in that work, and indeed the upper bounds on the rates of convergence implicit in that analysis may often have fairly complicated dependences on $\mathbb{C}$, Π_Θ, and $\mathcal{D}$, and furthermore often provide only very slow rates of convergence.

To derive explicit bounds on the rates of convergence, in the present work we specifically focus on families of *smooth* densities. The motivation for involving a notion of smoothness in characterizing rates of convergence is clear if we consider the extreme case in which Π_Θ contains two priors π_1 and π_2, with $\pi_1(\{h\}) = \pi_2(\{g\}) = 1$, where $\rho(h, g)$ is a very small but nonzero value; in this case, if we have only a small number of samples per task, we would require many tasks (on the order of $1/\rho(h, g)$) to observe any data points carrying any information that would

distinguish between these two priors (namely, points x with $h(x) \neq g(x)$); yet $\|\pi_1 - \pi_2\| = 1$, so that we have a slow rate of convergence (at least initially). A total boundedness condition on Π_Θ would limit the number of such pairs present in Π_Θ, so that for instance we cannot have arbitrarily close h and g, but less extreme variants of this can lead to slow asymptotic rates of convergence as well.

Specifically, in the present work we consider the following notion of smoothness. For $L \in (0, \infty)$ and $\alpha \in (0, 1]$, a function $f : \mathbb{C} \to \mathbb{R}$ is (L, α)-Hölder smooth if

$$\forall h, g \in \mathbb{C}, |f(h) - f(g)| \leq L\rho(h, g)^\alpha.$$

8.3 An Upper Bound

We now have the following theorem, holding for an arbitrary VC class $\mathbb{C}$ and data distribution $\mathcal{D}$; it is the main result of this work.

Theorem 8.1. *For Π_Θ any class of priors on $\mathbb{C}$ having (L, α)-Hölder smooth densities $\{f_\theta : \theta \in \Theta\}$, for any $T \in \mathbb{N}$, there exists an estimator $\hat{\theta}_{T\theta} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta), \ldots, \mathcal{Z}_{Td}(\theta))$ such that*

$$\sup_{\theta_\star \in \Theta} \mathbb{E}\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\| = \tilde{O}\left(LT^{-\frac{\alpha^2}{2(d+2\alpha)(\alpha+2(d+1))}}\right).$$

Proof. By the standard PAC analysis [Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989, Vapnik, 1982], for any $\gamma > 0$, with probability greater than $1-\gamma$, a sample of $k = O((d/\gamma)\log(1/\gamma))$ random points will partition $\mathbb{C}$ into regions of width less than γ. For brevity, we omit the t subscript on quantities such as $\mathcal{Z}_{tk}(\theta)$ throughout the following analysis, since the claims hold for any arbitrary value of t.

For any $\theta \in \Theta$, let π'_θ denote a (conditional on $X_1, \ldots, X_k$) distribution defined as follows. Let f'_θ denote the (conditional on $X_1, \ldots, X_k$) density function of π'_θ with respect to π_0, and for any $g \in \mathbb{C}$, let $f'_\theta(g) = \frac{\pi_\theta(\{h \in \mathbb{C} : \forall i \leq k, h(X_i)=g(X_i)\})}{\pi_0(\{h \in \mathbb{C} : \forall i \leq k, h(X_i)=g(X_i)\})}$ (or 0 if $\pi_0(\{h \in \mathbb{C} : \forall i \leq k, h(X_i) = g(X_i)\}) = 0$). In other words, π'_θ has the same probability mass as π_θ for each of the equivalence classes induced by $X_1, \ldots, X_k$, but conditioned on the equivalence class, simply has a constant-density

139

distribution over that equivalence class. Note that, by the smoothness condition, with probability greater than $1 - \gamma$, we have *everywhere*

$$|f_\theta(h) - f'_\theta(h)| < L\gamma^\alpha.$$

So for any $\theta, \theta' \in \Theta$, with probability greater than $1 - \gamma$,

$$\|\pi_\theta - \pi_{\theta'}\| = (1/2) \int |f_\theta - f_{\theta'}| \mathrm{d}\pi_0 < L\gamma^\alpha + (1/2) \int |f'_\theta - f'_{\theta'}| \mathrm{d}\pi_0.$$

Furthermore, since the regions that define f'_θ and $f'_{\theta'}$ are the same (namely, the partition induced by $X_1, \ldots, X_k$), we have

$$(1/2) \int |f'_\theta - f'_{\theta'}| \mathrm{d}\pi_0$$

$$= (1/2) \sum_{y_1, \ldots, y_k \in \{-1, +1\}} |\pi_\theta(\{h \in \mathbb{C} : \forall i \le k, h(X_i) = y_i\}) - \pi_{\theta'}(\{h \in \mathbb{C} : \forall i \le k, h(X_i) = y_i\})|$$

$$= \|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\|.$$

Thus, we have that with probability at least $1 - \gamma$,

$$\|\pi_\theta - \pi_{\theta'}\| < L\gamma^\alpha + \|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\|.$$

Following analogous to the inductive argument of Yang, Hanneke, and Carbonell [2011], suppose $I \subseteq \{1, \ldots, k\}$, fix $\bar{x}_I \in \mathcal{X}^{|I|}$ and $\bar{y}_I \in \{-1, +1\}^{|I|}$. Then the $\tilde{y}_I \in \{-1, +1\}^{|I|}$ for which no $h \in \mathbb{C}$ has $h(\bar{x}_I) = \tilde{y}_I$ for which $\|\bar{y}_I - \tilde{y}_I\|_1$ is minimal, has $\|\bar{y}_I - \tilde{y}_I\|_1 \le d + 1$, and for any $i \in I$ with $\bar{y}_i \ne \tilde{y}_i$, letting $\bar{y}'_j = \bar{y}_j$ for $j \in I \setminus \{i\}$ and $\bar{y}'_i = \tilde{y}_i$, we have

$$\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) = \mathbb{P}_{\mathbb{Y}_{I \setminus \{i\}}(\theta)|\mathbb{X}_{I \setminus \{i\}}}(\bar{y}_{I \setminus \{i\}}|\bar{x}_{I \setminus \{i\}}) - \mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I),$$

and similarly for θ', so that

$$|\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I)|$$

$$\le |\mathbb{P}_{\mathbb{Y}_{I \setminus \{i\}}(\theta)|\mathbb{X}_{I \setminus \{i\}}}(\bar{y}_{I \setminus \{i\}}|\bar{x}_{I \setminus \{i\}}) - \mathbb{P}_{\mathbb{Y}_{I \setminus \{i\}}(\theta')|\mathbb{X}_{I \setminus \{i\}}}(\bar{y}_{I \setminus \{i\}}|\bar{x}_{I \setminus \{i\}})|$$

$$+ |\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I)|.$$

Now consider that these two terms inductively define a binary tree. Every time the tree branches left once, it arrives at a difference of probabilities for a set I of one less element than that of its parent. Every time the tree branches right once, it arrives at a difference of probabilities for a $\bar{y}_I$ one closer to an unrealized $\tilde{y}_I$ than that of its parent. Say we stop branching the tree upon reaching a set I and a $\bar{y}_I$ such that either $\bar{y}_I$ is an unrealized labeling, or $|I| = d$. Thus, we can bound the original (root node) difference of probabilities by the sum of the differences of probabilities for the leaf nodes with $|I| = d$. Any path in the tree can branch left at most $k - d$ times (total) before reaching a set I with only d elements, and can branch right at most $d + 1$ times in a row before reaching a $\bar{y}_I$ such that both probabilities are zero, so that the difference is zero. So the depth of any leaf node with $|I| = d$ is at most $(k - d)d$. Furthermore, at any level of the tree, from left to right the nodes have strictly decreasing $|I|$ values, so that the maximum width of the tree is at most $k - d$. So the total number of leaf nodes with $|I| = d$ is at most $(k - d)^2 d$. Thus, for any $\bar{y} \in \{1, \ldots, k\}$ and $\bar{x} \in \mathcal{X}^k$,

$$
|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k}(\bar{y}|\bar{x}) - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}(\bar{y}|\bar{x})|
$$

$$
\leq (k - d)^2 d \cdot \max_{\bar{y}^d \in \{-1, +1\}^d} \max_{D \in \{1, \ldots, k\}^d} |\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\bar{y}^d|\bar{x}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\bar{y}^d|\bar{x}_D)|.
$$

Since

$$
\|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\| = (1/2) \sum_{\bar{y}^k \in \{-1, +1\}^k} |\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k}(\bar{y}^k) - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}(\bar{y}^k)|,
$$

and by Sauer's Lemma this is at most

$$
(ek)^d \max_{\bar{y}^k \in \{-1, +1\}^k} |\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k}(\bar{y}^k) - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}(\bar{y}^k)|,
$$

we have that

$$
\|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\| \leq (ek)^d k^2 d \max_{\bar{y}^d \in \{-1, +1\}^d} \max_{D \in \{1, \ldots, k\}^d} |\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_D}(\bar{y}^d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_D}(\bar{y}^d)|.
$$

141

Thus, we have that

$$\|\pi_\theta - \pi_{\theta'}\| = \mathbb{E}\|\pi_\theta - \pi_{\theta'}\|$$

$$< \gamma + L\gamma^\alpha + (ek)^d k^2 d\mathbb{E}\left[\max_{\bar{y}^d \in \{-1,+1\}^d} \max_{D \in \{1,\ldots,k\}^d} |\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_D}(\bar{y}^d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_D}(\bar{y}^d)|\right].$$

Note that

$$\mathbb{E}\left[\max_{\bar{y}^d \in \{-1,+1\}^d} \max_{D \in \{1,\ldots,k\}^d} |\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_D}(\bar{y}^d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_D}(\bar{y}^d)|\right]$$

$$\leq \sum_{\bar{y}^d \in \{-1,+1\}^d} \sum_{D \in \{1,\ldots,k\}^d} \mathbb{E}\left[|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_D}(\bar{y}^d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_D}(\bar{y}^d)|\right]$$

$$\leq (2k)^d \max_{\bar{y}^d \in \{-1,+1\}^d} \max_{D \in \{1,\ldots,k\}^d} \mathbb{E}\left[|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_D}(\bar{y}^d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_D}(\bar{y}^d)|\right],$$

and by exchangeability, this last line equals

$$(2k)^d \max_{\bar{y}^d \in \{-1,+1\}^d} \mathbb{E}\left[|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\bar{y}^d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\bar{y}^d)|\right].$$

Yang, Hanneke, and Carbonell [2011] showed that

$$\mathbb{E}\left[|\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\bar{y}^d) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\bar{y}^d)|\right] \leq 4\sqrt{\|\mathbb{P}_{\mathcal{Z}_d(\theta)} - \mathbb{P}_{\mathcal{Z}_d(\theta')}\|},$$

so that in total we have

$$\|\pi_\theta - \pi_{\theta'}\| < (L+1)\gamma^\alpha + 4(2ek)^{2d+2}\sqrt{\|\mathbb{P}_{\mathcal{Z}_d(\theta)} - \mathbb{P}_{\mathcal{Z}_d(\theta')}\|}.$$

Plugging in the value of $k = c(d/\gamma)\log(1/\gamma)$, this is

$$(L+1)\gamma^\alpha + 4\left(2ec\frac{d}{\gamma}\log\left(\frac{1}{\gamma}\right)\right)^{2d+2}\sqrt{\|\mathbb{P}_{\mathcal{Z}_d(\theta)} - \mathbb{P}_{\mathcal{Z}_d(\theta')}\|}.$$

So the only remaining question is the rate of convergence of our estimate of $\mathbb{P}_{\mathcal{Z}_d(\theta_\star)}$. If $N(\varepsilon)$ is the ε-covering number of $\{\mathbb{P}_{\mathcal{Z}_d(\theta)} : \theta \in \Theta\}$, then taking $\hat{\theta}_{T\theta_\star}$ as the minimum distance skeleton estimate of Devroye and Lugosi [2001], Yatracos [1985] achieves expected total variation distance ε from $\pi_{\theta_\star}$, for some $T = O((1/\varepsilon^2)\log N(\varepsilon/4))$. We can partition $\mathbb{C}$ into $O((L/\varepsilon)^{d/\alpha})$ cells of diameter $O((\varepsilon/L)^{1/\alpha})$, and set a constant density value within each cell, on an $O(\varepsilon)$-grid

of density values, and every prior with (L, α)-Hölder smooth density will have density within ε of some density so-constructed; there are then at most $(1/\varepsilon)^{O((L/\varepsilon)^{d/\alpha})}$ such densities, so this bounds the covering numbers of Π_Θ. Furthermore, the covering number of Π_Θ upper bounds $N(\varepsilon)$ [Yang, Hanneke, and Carbonell, 2011], so that $N(\varepsilon) \leq (1/\varepsilon)^{O((L/\varepsilon)^{d/\alpha})}$.

Solving $T = O(\varepsilon^{-2}(L/\varepsilon)^{d/\alpha} \log(1/\varepsilon))$ for ε, we have $\varepsilon = O\left(L\left(\frac{\log(TL)}{T}\right)^{\frac{\alpha}{d+2\alpha}}\right)$. So this bounds the rate of convergence for $\mathbb{E}\|\mathbb{P}_{\mathcal{Z}_d(\hat{\theta}_T)} - \mathbb{P}_{\mathcal{Z}_d(\theta_\star)}\|$, for $\hat{\theta}_T$ the minimum distance skeleton estimate. Plugging this rate into the bound on the priors, combined with Jensen's inequality, we have

$$\mathbb{E}\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\| < (L+1)\gamma^\alpha + 4\left(2ec\frac{d}{\gamma}\log\left(\frac{1}{\gamma}\right)\right)^{2d+2} O\left(L\left(\frac{\log(TL)}{T}\right)^{\frac{\alpha}{2d+4\alpha}}\right).$$

This holds for any $\gamma > 0$, so minimizing this expression over $\gamma > 0$ yields a bound on the rate. For instance, with $\gamma = \tilde{O}\left(T^{-\frac{\alpha}{2(d+2\alpha)(\alpha+2(d+1))}}\right)$, we have

$$\mathbb{E}\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\| = \tilde{O}\left(LT^{-\frac{\alpha^2}{2(d+2\alpha)(\alpha+2(d+1))}}\right).$$

$\square$

8.4 A Minimax Lower Bound

One natural quesiton is whether Theorem 8.1 can generally be improved. While we expect this to be true for some fixed VC classes (e.g., those of finite size), and in any case we expect that some of the constant factors in the exponent may be improvable, it is not at this time clear whether the general form of $T^{-\Theta(\alpha^2/(d+\alpha)^2)}$ is sometimes optimal. One way to investigate this question is to construct specific spaces $\mathbb{C}$ and distributions $\mathcal{D}$ for which a lower bound can be obtained. In particular, we are generally interested in exhibiting lower bounds that are worse than those that apply to the usual problem of density estimation based on direct access to the $h^*_{t\theta_\star}$ values (see Theorem 8.3 below).

Here we present a lower bound that is interesting for this reason. However, although larger than the optimal rate for methods wtih direct access to the target concepts, it is still far from

matching the upper bound above, so that the question of tightness remains open. Specifically, we have the following result.

Theorem 8.2. *For any integer $d \geq 1$, any $L > 0, \alpha \in (0,1]$, there is a value $C(d, L, \alpha) \in (0, \infty)$ such that, for any $T \in \mathbb{N}$, there exists an instance space $\mathcal{X}$, a concept space $\mathbb{C}$ of VC dimension d, a distribution $\mathcal{D}$ over $\mathcal{X}$, and a distribution π_0 over $\mathbb{C}$ such that, for Π_Θ a set of distributions over $\mathbb{C}$ with (L, α)-Hölder smooth density functions with respect to π_0, any estimator $\hat{\theta}_T = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta_\star), \dots, \mathcal{Z}_{Td}(\theta_\star))$ $(T = 1, 2, \dots)$, has*

$$\sup_{\theta_\star \in \Theta} \mathbb{E}\left[\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\|\right] \geq C(d, L, \alpha) T^{-\frac{\alpha}{2(d+\alpha)}}.$$

Proof. (Sketch) We proceed by a reduction from the task of determining the bias of a coin from among two given possibilities. Specifically, fix any $\gamma \in (0, 1/2), n \in \mathbb{N}$, and let $B_1(p), \dots, B_n(p)$ be i.i.d Bernoulli(p) random variables, for each $p \in [0, 1]$; then it is known that, for any (possibly nondeterministic) decision rule $\hat{p}_n : \{0, 1\}^n \to \{(1 + \gamma)/2, (1 - \gamma)/2\}$,

$$\frac{1}{2} \sum_{p \in \{(1+\gamma)/2, (1-\gamma)/2\}} \mathbb{P}(\hat{p}_n(B_1(p), \dots, B_n(p)) \neq p) \geq (1/32) \cdot \exp\left\{-128\gamma^2 n/3\right\}. \quad (8.1)$$

This easily follows from the results of Bar-Yossef [2003], Wald [1945], combined with a result of Poland and Hutter [2006] bounding the KL divergence.

To use this result, we construct a learning problem as follows. Fix some $m \in \mathbb{N}$ with $m \geq d$, let $\mathcal{X} = \{1, \dots, m\}$, and let $\mathbb{C}$ be the space of all classifiers $h : \mathcal{X} \to \{-1, +1\}$ such that $|\{x \in \mathcal{X} : h(x) = +1\}| \leq d$. Clearly the VC dimension of $\mathbb{C}$ is d. Define the distribution $\mathcal{D}$ as uniform over $\mathcal{X}$. Finally, we specify a family of (L, α)-Hölder smooth priors, parameterized by $\Theta = \{-1, +1\}^{\binom{m}{d}}$, as follows. Let $\gamma_m = (L/2)(1/m)^\alpha$. First, enumerate the $\binom{m}{d}$ distinct d-sized subsets of $\{1, \dots, m\}$ as $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_{\binom{m}{d}}$. Define the reference distribution π_0 by the property that, for any $h \in \mathbb{C}$, letting $q = |\{x : h(x) = +1\}|$, $\pi_0(\{h\}) = (\frac{1}{2})^d \binom{m-q}{d-q} / \binom{m}{d}$. For any $\mathbf{b} = (b_1, \dots, b_{\binom{m}{d}}) \in \{-1, 1\}^{\binom{m}{d}}$, define the prior $\pi_\mathbf{b}$ as the distribution of a random variable $h_\mathbf{b}$ specified by the following generative model. Let $i^* \sim \text{Uniform}(\{1, \dots, \binom{m}{d}\})$, let $C_\mathbf{b}(i^*) \sim \text{Bernoulli}((1 + \gamma_m b_{i^*})/2)$; finally, $h_\mathbf{b} \sim \text{Uniform}(\{h \in \mathbb{C} : \{x : h(x) = +1\} \subseteq$

144

$\mathcal{X}_{i^*}$, $\mathrm{Parity}(|\{x : h(x) = +1\}|) = C_{\mathbf{b}}(i^*)\})$, where $\mathrm{Parity}(n)$ is 1 if n is odd, or 0 if n is even. We will refer to the variables in this generative model below. For any $h \in \mathbb{C}$, letting $H = \{x : h(x) = +1\}$ and $q = |H|$, we can equivalently express $\pi_{\mathbf{b}}(\{h\}) = (\frac{1}{2})^d \binom{m}{d}^{-1} \sum_{i=1}^{\binom{m}{d}} \mathbb{1}[H \subseteq \mathcal{X}_i](1 + \gamma_m b_i)^{\mathrm{Parity}(q)}(1 - \gamma_m b_i)^{1-\mathrm{Parity}(q)}$. From this explicit representation, it is clear that, letting $f_{\mathbf{b}} = \frac{d\pi_{\mathbf{b}}}{d\pi_0}$, we have $f_{\mathbf{b}}(h) \in [1 - \gamma_m, 1 + \gamma_m]$ for all $h \in \mathbb{C}$. The fact that $f_{\mathbf{b}}$ is Hölder smooth follows from this, since every distinct $h, g \in \mathbb{C}$ have $\mathcal{D}(\{x : h(x) \neq g(x)\}) \geq 1/m = (2\gamma_m/L)^{1/\alpha}$.

Next we set up the reduction as follows. For any estimator $\hat{\pi}_T = \hat{\pi}_T(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star))$, and each $i \in \{1, \ldots, \binom{m}{d}\}$, let h_i be the classifier with $\{x : h_i(x) = +1\} = \mathcal{X}_i$; also, if $\hat{\pi}_T(\{h_i\}) > (\frac{1}{2})^d / \binom{m}{d}$, let $\hat{b}_i = 2\mathrm{Parity}(d) - 1$, and otherwise $\hat{b}_i = 1 - 2\mathrm{Parity}(d)$. We use these $\hat{b}_i$ values to estimate the original b_i values. Specifically, let $\hat{p}_i = (1 + \gamma_m \hat{b}_i)/2$ and $p_i = (1 + \gamma_m b_i)/2$, where $\mathbf{b} = \theta_\star$. Then

$$\|\hat{\pi}_T - \pi_{\theta_\star}\| \geq (1/2) \sum_{i=1}^{\binom{m}{d}} |\hat{\pi}_T(\{h_i\}) - \pi_{\theta_\star}(\{h_i\})|$$

$$\geq (1/2) \sum_{i=1}^{\binom{m}{d}} \frac{\gamma_m}{2^d \binom{m}{d}} |\hat{b}_i - b_i|/2 = (1/2) \sum_{i=1}^{\binom{m}{d}} \frac{1}{2^d \binom{m}{d}} |\hat{p}_i - p_i|.$$

Thus, we have reduced from the problem of deciding the biases of these $\binom{m}{d}$ independent Bernoulli random variables. To complete the proof, it suffices to lower bound the expectation of the right side for an *arbitrary* estimator.

Toward this end, we in fact study an even easier problem. Specifically, consider an estimator $\hat{q}_i = \hat{q}_i(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star), i_1^*, \ldots, i_T^*)$, where i_t^* is the i^* random variable in the generative model that defines $h_{t\theta_\star}^*$; that is, $i_t^* \sim \mathrm{Uniform}(\{1, \ldots, \binom{m}{d}\})$, $C_t \sim \mathrm{Bernoulli}((1 + \gamma_m b_{i_t^*})/2)$, and $h_{t\theta_\star}^* \sim \mathrm{Uniform}(\{h \in \mathbb{C} : \{x : h(x) = +1\} \subseteq \mathcal{X}_{i_t^*}, \mathrm{Parity}(|\{x : h(x) = +1\}|) = C_t\})$, where the i_t^* are independent across t, as are the C_t and $h_{t\theta_\star}^*$. Clearly the $\hat{p}_i$ from above can be viewed as an estimator of this type, which simply ignores the knowledge of i_t^*. The knowledge of these i_t^* variables simplifies the analysis, since given $\{i_t^* : t \leq T\}$, the data can be partitioned into $\binom{m}{d}$ disjoint sets, $\{\{\mathcal{Z}_{td}(\theta_\star) : i_t^* = i\} : i = 1, \ldots, \binom{m}{d}\}$, and we can

use only the set $\{\mathcal{Z}_{td}(\theta_\star) : i_t^* = i\}$ to estimate p_i. Furthermore, we can use only the subset of these for which $\mathbb{X}_{td} = \mathcal{X}_i$, since otherwise we have zero information about the value of $\mathrm{Parity}(|\{x : h_{t\theta_\star}^*(x) = +1\}|)$. That is, given $i_t^* = i$, any $\mathcal{Z}_{td}(\theta_\star)$ is conditionally independent from every b_j for $j \neq i$, and is even conditionally independent from b_i when $\mathbb{X}_{td}$ is not completely contained in $\mathcal{X}_i$; specifically, in this case, regardless of b_i, the conditional distribution of $\mathbb{Y}_{td}(\theta_\star)$ given $i_t^* = i$ and given $\mathbb{X}_{td}$ is a product distribution, which deterministically assigns label -1 to those $Y_{tk}(\theta_\star)$ with $X_{tk} \notin \mathcal{X}_i$, and gives uniform random values to the subset of $\mathbb{Y}_{td}(\theta_\star)$ with their respective $X_{tk} \in \mathcal{X}_i$. Finally, letting $r_t = \mathrm{Parity}(|\{k \leq d : Y_{tk}(\theta_\star) = +1\}|)$, we note that given $i_t^* = i$, $\mathbb{X}_{td} = \mathcal{X}_i$, and the value r_t, b_i is conditionally independent from $\mathcal{Z}_{td}(\theta_\star)$. Thus, the set of values $C_{iT}(\theta_\star) = \{r_t : i_t^* = i, \mathbb{X}_{td} = \mathcal{X}_i\}$ is a sufficient statistic for b_i (hence for p_i). Recall that, when $i_t^* = i$ and $\mathbb{X}_{td} = \mathcal{X}_i$, the value of r_t is equal to C_t, a $\mathrm{Bernoulli}(p_i)$ random variable. Thus, we neither lose nor gain anything (in terms of risk) by restricting ourselves to estimators $\hat{q}_i$ of the type $\hat{q}_i = \hat{q}_i(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star), i_1^*, \ldots, i_T^*) = \hat{q}_i'(C_{iT}(\theta_\star))$, for some $\hat{q}_i'$ [Schervish, 1995]: that is, estimators that are a function of the $N_{iT}(\theta_\star) = |C_{iT}(\theta_\star)|$ $\mathrm{Bernoulli}(p_i)$ random variables, which we should note are conditionally i.i.d. given $N_{iT}(\theta_\star)$.

Thus, by (8.1), for any $n \leq T$,

$$\frac{1}{2} \sum_{b_i \in \{-1,+1\}} \mathbb{E}\left[|\hat{q}_i - p_i| \Big| N_{iT}(\theta_\star) = n\right] = \frac{1}{2} \sum_{b_i \in \{-1,+1\}} \gamma_m \mathbb{P}\left(\hat{q}_i \neq p_i \Big| N_{iT}(\theta_\star) = n\right)$$

$$\geq (\gamma_m/32) \cdot \exp\left\{-128\gamma_m^2 N_i/3\right\}.$$

Also note that, for each i, $\mathbb{E}[N_i] = \frac{d!(1/m)^d}{\binom{m}{d}}T \leq (d/m)^{2d}T = d^{2d}(2\gamma_m/L)^{2d/\alpha}T$, so that Jensen's inequality, linearity of expectation, and the law of total expectation imply

$$\frac{1}{2} \sum_{b_i \in \{-1,+1\}} \mathbb{E}[|\hat{q}_i - p_i|] \geq (\gamma_m/32) \cdot \exp\left\{-43(2/L)^{2d/\alpha}d^{2d}\gamma_m^{2+2d/\alpha}T\right\}.$$

Thus, by linearity of the expectation,

$$\left(\frac{1}{2}\right)^{\binom{m}{d}} \sum_{\mathbf{b}\in\{-1,+1\}^{\binom{m}{d}}} \mathbb{E}\left[\sum_{i=1}^{\binom{m}{d}} \frac{1}{2^d\binom{m}{d}}|\hat{q}_i - p_i|\right] = \sum_{i=1}^{\binom{m}{d}} \frac{1}{2^d\binom{m}{d}} \frac{1}{2} \sum_{b_i\in\{-1,+1\}} \mathbb{E}\left[|\hat{q}_i - p_i|\right]$$

$$\geq (\gamma_m/(32 \cdot 2^d)) \cdot \exp\left\{-43(2/L)^{2d/\alpha}d^{2d}\gamma_m^{2+2d/\alpha}T\right\}.$$

In particular, taking

$$m = \left\lceil (L/2)^{1/\alpha} \left(\frac{T}{43(2/L)^{2d/\alpha}d^{2d}}\right)^{\frac{1}{2(d+\alpha)}} \right\rceil,$$

we have

$$\gamma_m = \Theta\left(\left(\frac{43(2/L)^{2d/\alpha}d^{2d}}{T}\right)^{\frac{\alpha}{2(d+\alpha)}}\right),$$

so that

$$\left(\frac{1}{2}\right)^{\binom{m}{d}} \sum_{\mathbf{b}\in\{-1,+1\}^{\binom{m}{d}}} \mathbb{E}\left[\sum_{i=1}^{\binom{m}{d}} \frac{1}{2^d\binom{m}{d}}|\hat{q}_i - p_i|\right] = \Omega\left(2^{-d}\left(\frac{43(2/L)^{2d/\alpha}d^{2d}}{T}\right)^{\frac{\alpha}{2(d+\alpha)}}\right).$$

In particular, this implies there exists some **b** for which

$$\mathbb{E}\left[\sum_{i=1}^{\binom{m}{d}} \frac{1}{2^d\binom{m}{d}}|\hat{q}_i - p_i|\right] = \Omega\left(2^{-d}\left(\frac{43(2/L)^{2d/\alpha}d^{2d}}{T}\right)^{\frac{\alpha}{2(d+\alpha)}}\right).$$

Applying this lower bound to the estimator $\hat{p}_i$ defined above yields the result. $\qquad\square$

In the extreme case of allowing arbitrary dependence on the data samples, we merely recover the known results lower bounding the risk of density estimation from i.i.d. samples from a smooth density, as indicated by the following result.

Theorem 8.3. *For any integer $d \geq 1$, there exists an instance space $\mathcal{X}$, a concept space $\mathbb{C}$ of VC dimension d, a distribution $\mathcal{D}$ over $\mathcal{X}$, and a distribution π_0 over $\mathbb{C}$ such that, for Π_Θ the set of distributions over $\mathbb{C}$ with (L,α)-Hölder smooth density functions with respect to π_0, any sequence of estimators, $\hat{\theta}_T = \hat{\theta}_T(\mathcal{Z}_1(\theta_\star), \ldots, \mathcal{Z}_T(\theta_\star))$ ($T = 1, 2, \ldots$), has*

$$\sup_{\theta_\star\in\Theta} \mathbb{E}\left[\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\|\right] = \Omega\left(T^{-\frac{\alpha}{d+2\alpha}}\right).$$

The proof is a simple reduction from the problem of estimating $\pi_{\theta_\star}$ based on direct access to $h_{1\theta_\star}^*, \ldots, h_{T\theta_\star}^*$, which is essentially equivalent to the standard model of density estimation, and indeed the lower bound in Theorem 8.3 is a well-known result for density estimation from T i.i.d. samples from a Hölder smooth density in a d-dimensional space [see e.g., Devroye and Lugosi, 2001].

8.5   Future Directions

There are several interesting questions that remain open at this time. Can either the lower bound or upper bound be improved in general? If, instead of d samples per task, we instead use $m \geq d$ samples, how does the minimax risk vary with m? Related to this, what is the optimal value of m to optimize the rate of convergence as a function of mT, the total number of samples? More generally, if an estimator is permitted to use N total samples, taken from however many tasks it wishes, what is the optimal rate of convergence as a function of N?

Chapter 9

Estimation of Priors with Applications to Preference Elicitation

Abstract

[1]We extend the work of [Yang, Hanneke, and Carbonell, 2013] on estimating prior distributions over VC classes to the case of real-valued functions in a VC subgraph class. We then apply this technique to the problem of maximizing customer satisfaction using a minimal number of value queries in an online preference elicitation scenario.

9.1   Introduction

Consider an online travel agency, where customers go to the site with some idea of what type of travel they are interested in; the site then poses a series of questions to each customer, and identifies a travel package that best suits their desires, budget, and dates. There are many options of travel packages, with options on location, site-seeing tours, hotel and room quality, etc. Because of this, serving the needs of an *arbitrary* customer might be a lengthy process, requiring many

---

[1]This chapter is based on joint work with Steve Hanneke

detailed questions. Fortunately, the stream of customers is typically not a worst-case sequence, and in particular obeys many statistical regularities: in particular, it is not too far from reality to think of the customers as being independent and identically distributed samples. With this assumption in mind, it becomes desirable to identify some of these statistical regularities so that we can pose the questions that are typically most relevant, and thereby more quickly identify the travel package that best suits the needs of the typical customer. One straightforward way to do this is to directly *estimate* the distribution of customer value functions, and optimize the questioning system to minimize the expected number of questions needed to find a suitable travel package.

One can model this problem in the style of Bayesian combinatorial auctions, in which each customer has a value function for each possible bundle of items. However, it is slightly different, in that we do not assume the distribution of customers is known, but rather are interested in estimating this distribution; the obtained estimate can then be used in combination with methods based on Bayesian decision theory. In contrast to the literature on Bayesian auctions (and subjectivist Bayesian decision theory in general), this technique is able to maintain general guarantees on performance that hold under an objective interpretation of the problem, rather than merely guarantees holding under an arbitrary assumed prior belief. This general idea is sometimes referred to as *Empirical Bayesian* decision theory in the machine learning and statistics literatures. The ideal result for an Empirical Bayesian algorithm is to be competitive with the corresponding Bayesian methods based on the *actual* distribution of the data (assuming the data are random, with an unknown distribution); that is, although the Empirical Bayesian methods only operate with a data-based estimate of the distribution, the aim is to perform nearly as well as methods based on the true (unobservable) distribution. In this work, we present results of this type, in the context of an abstraction of the aforementioned online travel agency problem, where the measure of performance is the expected number of questions to find a suitable package.

The technique we use here is rooted in the work of [Yang, Hanneke, and Carbonell, 2013] on

transfer learning with a VC class. The component of that work of interest here is the estimation of prior distributions over VC classes. Essentially, there is a given class of functions, from which a sequence of functions is sampled i.i.d. according to an unknown distribution. We observe a number of values of each of these functions, evaluated at points chosen at random, and are then tasked with estimating the distribution of these functions. This is more challenging than the traditional problem of nonparametric density estimation, since we are not permitted direct access to these functions, but rather only a limited number of evaluations of the function (i.e., a number of $(x, f(x))$ pairs). The work of [Yang, Hanneke, and Carbonell, 2013] develops a technique for estimating the distribution of these functions, given that the functions are binary-valued, the class of functions has finite VC dimension, and the class of distributions is totally bounded. In this work, we extend this technique to classes of real-valued functions having finite pseudo-dimension, a natural generalization of VC dimension for real-valued functions [Haussler, 1992].

The specific application we are interested in here may be expressed abstractly as a kind of combinatorial auction with preference elicitation. Specifically, we suppose there is a collection of items on a menu, and each possible bundle of items has an associated fixed price. There is a stream of customers, each with a valuation function that provides a value for each possible bundle of items. The objective is to serve each customer a bundle of items that nearly-maximizes his or her surplus value (value minus price). However, we are not permitted direct observation of the customer valuation functions; rather, we may query for the value of any given bundle of items; this is referred to as a *value query* in the literature on preference elicitation in combinatorial auctions (see Chapter 14 of [Cramton, Shoham, and Steinberg, 2006], [Zinkevich, Blum, and Sandholm, 2003]). The objective is to achieve this near-maximal surplus guarantee, while making only a small number of queries per customer. We suppose the customer valuation function are sampled i.i.d. according to an unknown distribution over a known (but arbitrary) class of real-valued functions having finite pseudo-dimension. Reasoning that knowledge of this distribution

should allow one to make a smaller number of value queries per customer, we are interested in estimating this unknown distribution, so that as we serve more and more customers, the number of queries per customer required to identify a near-optimal bundle should decrease. In this context, we in fact prove that in the limit, the expected number of queries per customer converges to the number required of a method having direct knowledge of the true distribution of valuation functions.

9.2 Notation

Let $\mathcal{B}$ denote a σ-algebra on $\mathcal{X} \times \mathbb{R}$, let $\mathcal{B}_\mathcal{X}$ denote the σ-algebra on $\mathcal{X}$. Also let $\rho(h,g) = \int |h - g| \mathrm{d}P_X$, where P_X is a marginal distribution over $\mathcal{X}$. Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$ with Borel σ-algebra $\mathcal{B}_\mathcal{F}$ induced by ρ. Let Θ be a set of parameters, and for each $\theta \in \Theta$, let π_θ denote a probability measure on $(\mathcal{F}, \mathcal{B}_\mathcal{F})$. We suppose $\{\pi_\theta : \theta \in \Theta\}$ is totally bounded in total variation distance, and that $\mathcal{F}$ is a uniformly bounded VC subgraph class with pseudodimension d. We also suppose ρ is a *metric* when restricted to $\mathcal{F}$.

Let $\{X_{ti}\}_{t,i \in \mathbb{N}}$ be i.i.d. P_X random variables. For each $\theta \in \Theta$, let $\{h_{t\theta}^*\}_{t \in \mathbb{N}}$ be i.i.d. π_θ random variables, independent from $\{X_{ti}\}_{t,i \in \mathbb{N}}$. For each $t \in \mathbb{N}$ and $\theta \in \Theta$, let $Y_{ti}(\theta) = h_{t\theta}^*(X_{ti})$ for $i \in \mathbb{N}$, and let $\mathcal{Z}_t(\theta) = \{(X_{t1}, Y_{t1}(\theta)), (X_{t2}, Y_{t2}(\theta)), \ldots\}$, $\mathbb{X}_t = \{X_{t1}, X_{t2}, \ldots\}$, and $\mathbb{Y}_t(\theta) = \{Y_{t1}(\theta), Y_{t2}(\theta), \ldots\}$; for each $k \in \mathbb{N}$, define $\mathcal{Z}_{tk}(\theta) = \{(X_{t1}, Y_{t1}(\theta)), \ldots, (X_{tk}, Y_{tk}(\theta))\}$, $\mathbb{X}_{tk} = \{X_{t1}, \ldots, X_{tk}\}$, and $\mathbb{Y}_{tk}(\theta) = \{Y_{t1}(\theta), \ldots, Y_{tk}(\theta)\}$.

For any probability measures μ, μ', we denote the total variation distance by

$$\|\mu - \mu'\| = \sup_A \mu(A) - \mu'(A),$$

where A ranges over measurable sets.

Lemma 9.1. *For any $\theta, \theta' \in \Theta$ and $t \in \mathbb{N}$,*

$$\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\|.$$

152

Proof. Fix $\theta, \theta' \in \Theta$, $t \in \mathbb{N}$. Let $\mathbb{X} = \{X_{t1}, X_{t2}, \ldots\}$, $\mathbb{Y}(\theta) = \{Y_{t1}(\theta), Y_{t2}(\theta), \ldots\}$, and for $k \in \mathbb{N}$ let $\mathbb{X}_k = \{X_{t1}, \ldots, X_{tk}\}$. and $\mathbb{Y}_k(\theta) = \{Y_{t1}(\theta), \ldots, Y_{tk}(\theta)\}$. For $h \in \mathcal{F}$, let $c_{\mathbb{X}}(h) = \{(X_{t1}, h(X_{t1})), (X_{t2}, h(X_{t2})), \ldots\}$.

For $h, g \in \mathcal{F}$, define $\rho_{\mathbb{X}}(h, g) = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} |h(X_{ti}) - g(X_{ti})|$ (if the limit exists), and $\rho_{\mathbb{X}_k}(h, g) = \frac{1}{k} \sum_{i=1}^{k} |h(X_{ti}) - g(X_{ti})|$. Note that since $\mathcal{F}$ is a uniformly bounded VC subgraph class, so is the collection of functions $\{|h - g| : h, g \in \mathcal{F}\}$, so that the uniform strong law of large numbers implies that with probability one, $\forall h, g \in \mathcal{F}$, $\rho_{\mathbb{X}}(h, g)$ exists and has $\rho_{\mathbb{X}}(h, g) = \rho(h, g)$ [Vapnik, 1982].

Consider any $\theta, \theta' \in \Theta$, and any $A \in \mathcal{B}_{\mathcal{F}}$. Then any $h \notin A$ has $\forall g \in A$, $\rho(h, g) > 0$ (by the metric assumption). Thus, if $\rho_{\mathbb{X}}(h, g) = \rho(h, g)$ for all $h, g \in \mathcal{F}$, then $\forall h \notin A$,

$$\forall g \in A, \rho_{\mathbb{X}}(h, g) = \rho(h, g) > 0 \implies \forall g \in A, c_{\mathbb{X}}(h) \neq c_{\mathbb{X}}(g) \implies c_{\mathbb{X}}(h) \notin c_{\mathbb{X}}(A).$$

This implies $c_{\mathbb{X}}^{-1}(c_{\mathbb{X}}(A)) = A$. Under these conditions,

$$\mathbb{P}_{\mathcal{Z}_t(\theta)|\mathbb{X}}(c_{\mathbb{X}}(A)) = \pi_\theta(c_{\mathbb{X}}^{-1}(c_{\mathbb{X}}(A))) = \pi_\theta(A),$$

and similarly for θ'.

Any measurable set C for the range of $\mathcal{Z}_t(\theta)$ can be expressed as $C = \{c_{\bar{x}}(h) : (h, \bar{x}) \in C'\}$ for some appropriate $C' \in \mathcal{B}_{\mathcal{F}} \otimes \mathcal{B}_{\mathcal{X}}^{\infty}$. Letting $C'_{\bar{x}} = \{h : (h, \bar{x}) \in C'\}$, we have

$$\mathbb{P}_{\mathcal{Z}_t(\theta)}(C) = \int \pi_\theta(c_{\bar{x}}^{-1}(c_{\bar{x}}(C'_{\bar{x}}))) \mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) = \int \pi_\theta(C'_{\bar{x}}) \mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) = \mathbb{P}_{(h_{t\theta}^*, \mathbb{X})}(C').$$

Likewise, this reasoning holds for θ'. Then

$$\|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| = \|\mathbb{P}_{(h_{t\theta}^*, \mathbb{X})} - \mathbb{P}_{(h_{t\theta'}^*, \mathbb{X})}\|$$

$$= \sup_{C' \in \mathcal{B}_{\mathcal{F}} \otimes \mathcal{B}_{\mathcal{X}}^{\infty}} \left| \int (\pi_\theta(C'_{\bar{x}}) - \pi_{\theta'}(C'_{\bar{x}})) \mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) \right|$$

$$\leq \int \sup_{A \in \mathcal{B}_{\mathcal{F}}} |\pi_\theta(A) - \pi_{\theta'}(A)| \mathbb{P}_{\mathbb{X}}(\mathrm{d}\bar{x}) = \|\pi_\theta - \pi_{\theta'}\|.$$

Since $h_{t\theta}^*$ and $\mathbb{X}$ are independent, for $A \in \mathcal{B}_{\mathcal{F}}$, $\pi_\theta(A) = \mathbb{P}_{h_{t\theta}^*}(A) = \mathbb{P}_{h_{t\theta}^*}(A) \mathbb{P}_{\mathbb{X}}(\mathcal{X}^\infty) =$

153

$\mathbb{P}_{(h_{t\theta}^*, \mathbb{X})}(A \times \mathcal{X}^\infty)$. Analogous reasoning holds for $h_{t\theta'}^*$. Thus, we have

$$\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{(h_{t\theta}^*, \mathbb{X})}(\cdot \times \mathcal{X}^\infty) - \mathbb{P}_{(h_{t\theta'}^*, \mathbb{X})}(\cdot \times \mathcal{X}^\infty)\|$$

$$\leq \|\mathbb{P}_{(h_{t\theta}^*, \mathbb{X})} - \mathbb{P}_{(h_{t\theta'}^*, \mathbb{X})}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\|.$$

Combining the above, we have $\|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| = \|\pi_\theta - \pi_{\theta'}\|$. $\qquad\square$

Lemma 9.2. *There exists a sequence $r_k = o(1)$ such that, $\forall t, k \in \mathbb{N}$, $\forall \theta, \theta' \in \Theta$,*

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq \|\pi_\theta - \pi_{\theta'}\| \leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + r_k.$$

Proof. This proof follows identically to a proof of [Yang, Hanneke, and Carbonell, 2013], but is included here for completeness. Since $\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(A) = \mathbb{P}_{\mathcal{Z}_t(\theta)}(A \times (\mathcal{X} \times \mathbb{R})^\infty)$ for all measurable $A \subseteq (\mathcal{X} \times \mathbb{R})^k$, and similarly for θ', we have

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| = \sup_{A \in \mathcal{B}^k} \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(A) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(A)$$

$$= \sup_{A \in \mathcal{B}^k} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A \times (\mathcal{X} \times \mathbb{R})^\infty) - \mathbb{P}_{\mathcal{Z}_t(\theta')}(A \times (\mathcal{X} \times \mathbb{R})^\infty)$$

$$\leq \sup_{A \in \mathcal{B}^\infty} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A) - \mathbb{P}_{\mathcal{Z}_t(\theta')}(A) = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\|,$$

which implies the left inequality when combined with Lemma 9.1.

Next, we focus on the right inequality. Fix $\theta, \theta' \in \Theta$ and $\gamma > 0$, and let $B \in \mathcal{B}^\infty$ be such that

$$\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| < \mathbb{P}_{\mathcal{Z}_t(\theta)}(B) - \mathbb{P}_{\mathcal{Z}_t(\theta')}(B) + \gamma.$$

Let $\mathcal{A} = \{A \times (\mathcal{X} \times \mathbb{R})^\infty : A \in \mathcal{B}^k, k \in \mathbb{N}\}$. Note that $\mathcal{A}$ is an algebra that generates $\mathcal{B}^\infty$. Thus, Carathéodory's extension theorem [Schervish, 1995] implies that there exist disjoint sets $\{A_i\}_{i \in \mathbb{N}}$ in $\mathcal{A}$ such that $B \subseteq \bigcup_{i \in \mathbb{N}} A_i$ and

$$\mathbb{P}_{\mathcal{Z}_t(\theta)}(B) - \mathbb{P}_{\mathcal{Z}_t(\theta')}(B) < \sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i) + \gamma.$$

Since these A_i sets are disjoint, each of these sums is bounded by a probability value, which implies that there exists some $n \in \mathbb{N}$ such that

$$\sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) < \gamma + \sum_{i=1}^{n} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i),$$

154

which implies

$$\sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i \in \mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i) < \gamma + \sum_{i=1}^{n} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i=1}^{n} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i)$$

$$= \gamma + \mathbb{P}_{\mathcal{Z}_t(\theta)} \left(\bigcup_{i=1}^{n} A_i \right) - \mathbb{P}_{\mathcal{Z}_t(\theta')} \left(\bigcup_{i=1}^{n} A_i \right).$$

As $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$, there exists $m \in \mathbb{N}$ and measurable $B_m \in \mathcal{B}^m$ such that $\bigcup_{i=1}^{n} A_i = B_m \times (\mathcal{X} \times \mathbb{R})^{\infty}$, and therefore

$$\mathbb{P}_{\mathcal{Z}_t(\theta)} \left(\bigcup_{i=1}^{n} A_i \right) - \mathbb{P}_{\mathcal{Z}_t(\theta')} \left(\bigcup_{i=1}^{n} A_i \right) = \mathbb{P}_{\mathcal{Z}_{tm}(\theta)}(B_m) - \mathbb{P}_{\mathcal{Z}_{tm}(\theta')}(B_m)$$

$$\leq \|\mathbb{P}_{\mathcal{Z}_{tm}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tm}(\theta')}\| \leq \lim_{k \to \infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.$$

Combining the above, we have $\|\pi_\theta - \pi_{\theta'}\| \leq \lim_{k \to \infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + 3\gamma$. By letting γ approach 0, we have

$$\|\pi_\theta - \pi_{\theta'}\| \leq \lim_{k \to \infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.$$

So there exists a sequence $r_k(\theta, \theta') = o(1)$ such that

$$\forall k \in \mathbb{N}, \|\pi_\theta - \pi_{\theta'}\| \leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + r_k(\theta, \theta').$$

Now let $\gamma > 0$ and let Θ_γ be a minimal γ-cover of Θ. Define the quantity $r_k(\gamma) = \max_{\theta, \theta' \in \Theta_\gamma} r_k(\theta, \theta')$. Then for any $\theta, \theta' \in \Theta$, let $\theta_\gamma = \operatorname{argmin}_{\theta'' \in \Theta_\gamma} \|\pi_\theta - \pi_{\theta''}\|$ and $\theta'_\gamma = \operatorname{argmin}_{\theta'' \in \Theta_\gamma} \|\pi_{\theta'} - \pi_{\theta''}\|$. Then a triangle inequality implies that $\forall k \in \mathbb{N}$,

$$\|\pi_\theta - \pi_{\theta'}\| \leq \|\pi_\theta - \pi_{\theta_\gamma}\| + \|\pi_{\theta_\gamma} - \pi_{\theta'_\gamma}\| + \|\pi_{\theta'_\gamma} - \pi_{\theta'}\|$$

$$< 2\gamma + r_k(\theta_\gamma, \theta'_\gamma) + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\|$$

$$\leq 2\gamma + r_k(\gamma) + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\|.$$

Triangle inequalities and the left inequality from the lemma statement (already established) imply

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\|$$

$$\leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|$$

$$\leq \|\pi_{\theta_\gamma} - \pi_\theta\| + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + \|\pi_{\theta'_\gamma} - \pi_{\theta'}\|$$

$$< 2\gamma + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.$$

So in total we have

$$\|\pi_\theta - \pi_{\theta'}\| \leq 4\gamma + r_k(\gamma) + \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.$$

Since this holds for all $\gamma > 0$, defining $r_k = \inf_{\gamma>0}(4\gamma + r_k(\gamma))$, we have the right inequality of the lemma statement. Furthermore, since each $r_k(\theta, \theta') = o(1)$, and $|\Theta_\gamma| < \infty$, we have $r_k(\gamma) = o(1)$ for each $\gamma > 0$, and thus we also have $r_k = o(1)$. $\qquad\square$

Lemma 9.3. $\forall t, k \in \mathbb{N}$, there exists a monotone function $M_k(x) = o(1)$ such that, $\forall \theta, \theta' \in \Theta$,

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq M_k\left(\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|\right).$$

Proof. Fix any $t \in \mathbb{N}$, and let $\mathbb{X} = \{X_{t1}, X_{t2}, \ldots\}$ and $\mathbb{Y}(\theta) = \{Y_{t1}(\theta), Y_{t2}(\theta), \ldots\}$, and for $k \in \mathbb{N}$ let $\mathbb{X}_k = \{X_{t1}, \ldots, X_{tk}\}$ and $\mathbb{Y}_k(\theta) = \{Y_{t1}(\theta), \ldots, Y_{tk}(\theta)\}$.

If $k \leq d$, then $\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(\cdot) = \mathbb{P}_{\mathcal{Z}_{td}(\theta)}(\cdot \times (\mathcal{X} \times \{-1, +1\})^{d-k})$, so that

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \leq \|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|,$$

and therefore the result trivially holds.

Now suppose $k > d$. Fix any $\gamma > 0$, and let $B_{\theta,\theta'} \subseteq (\mathcal{X} \times \mathbb{R})^k$ be a measurable set such that

$$\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(B_{\theta,\theta'}) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(B_{\theta,\theta'}) \leq \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|$$

$$\leq \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(B_{\theta,\theta'}) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(B_{\theta,\theta'}) + \gamma.$$

By Carathéodory's extension theorem, there exists a disjoint sequence of sets $\{B_i\}_{i=1}^\infty$ such that

$$\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(B_{\theta,\theta'}) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(B_{\theta,\theta'}) < \gamma + \sum_{i=1}^\infty \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(B_i) - \sum_{i=1}^\infty \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(B_i),$$

156

and such that each $B_i(\theta, \theta')$ is representable as follows; for some $\ell_i(\theta, \theta') \in \mathbb{N}$, and sets $C_{ij} = (A_{ij1} \times (-\infty, t_{ij1}]) \times \cdots \times (A_{ijk} \times (-\infty, t_{ijk}])$, for $j \leq \ell_i(\theta, \theta')$, where each $A_{ijp} \in \mathcal{B}_\mathcal{X}$, the set $B_i(\theta, \theta')$ is representable as $\bigcup_{s \in S_i} \bigcap_{j=1}^{\ell_i(\theta, \theta')} D_{ijs}$, where $S_i \subseteq \{0, \ldots, 2^{\ell_i(\theta, \theta')} - 1\}$, each $D_{ijs} \in \{C_{ij}, C_{ij}^c\}$, and $s \neq s' \Rightarrow \bigcap_{j=1}^{\ell_i(\theta, \theta')} D_{ijs} \cap \bigcap_{j=1}^{\ell_i(\theta, \theta')} D_{ijs'} = \emptyset$. Since the $B_i(\theta, \theta')$ are disjoint, the above sums are bounded, so that there exists $m_k(\theta, \theta', \gamma) \in \mathbb{N}$ such that every $m \geq m_k(\theta, \theta', \gamma)$ has

$$\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(B_{\theta, \theta'}) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(B_{\theta, \theta'})$$
$$< 2\gamma + \sum_{i=1}^{m} \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(B_i(\theta, \theta')) - \sum_{i=1}^{m} \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(B_i(\theta, \theta')),$$

Now define $\tilde{M}_k(\gamma) = \max_{\theta, \theta' \in \Theta_\gamma} m_k(\theta, \theta', \gamma)$. Then for any $\theta, \theta' \in \Theta$, let $\theta_\gamma, \theta'_\gamma \in \Theta_\gamma$ be such that $\|\pi_\theta - \pi_{\theta_\gamma}\| < \gamma$ and $\|\pi_{\theta'} - \pi_{\theta'_\gamma}\| < \gamma$, which implies $\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)}\| < \gamma$ and $\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta')} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\| < \gamma$ by Lemma 9.2. Then

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| < \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\| + 2\gamma$$
$$\leq \mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)}(B_{\theta_\gamma, \theta'_\gamma}) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}(B_{\theta_\gamma, \theta'_\gamma}) + 3\gamma$$
$$\leq \sum_{i=1}^{\tilde{M}_k(\gamma)} \mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)}(B_i(\theta_\gamma, \theta'_\gamma)) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}(B_i(\theta_\gamma, \theta'_\gamma)) + 5\gamma.$$

Again, since the $B_i(\theta_\gamma, \theta'_\gamma)$ are disjoint, this equals

$$5\gamma + \mathbb{P}_{\mathcal{Z}_{tk}(\theta_\gamma)}\left(\bigcup_{i=1}^{\tilde{M}_k(\gamma)} B_i(\theta_\gamma, \theta'_\gamma)\right) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta'_\gamma)}\left(\bigcup_{i=1}^{\tilde{M}_k(\gamma)} B_i(\theta_\gamma, \theta'_\gamma)\right)$$
$$\leq 7\gamma + \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}\left(\bigcup_{i=1}^{\tilde{M}_k(\gamma)} B_i(\theta_\gamma, \theta'_\gamma)\right) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\left(\bigcup_{i=1}^{\tilde{M}_k(\gamma)} B_i(\theta_\gamma, \theta'_\gamma)\right)$$
$$= 7\gamma + \sum_{i=1}^{\tilde{M}_k(\gamma)} \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(B_i(\theta_\gamma, \theta'_\gamma)) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(B_i(\theta_\gamma, \theta'_\gamma))$$
$$\leq 7\gamma + \tilde{M}_k(\gamma) \max_{i \leq \tilde{M}_k(\gamma)} \left|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(B_i(\theta_\gamma, \theta'_\gamma)) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(B_i(\theta_\gamma, \theta'_\gamma))\right|.$$

Thus, if we can show that each $\left|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(B_i(\theta_\gamma, \theta'_\gamma)) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(B_i(\theta_\gamma, \theta'_\gamma))\right|$ is bounded by a $o(1)$

157

function of $\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|$, then the result will follow by substituting this relaxation into the above expression and defining M_k by minimizing the resulting expression over $\gamma > 0$.

Toward this end, let C_{ij} be as above from the definition of $B_i(\theta_\gamma, \theta'_\gamma)$, and note that $I_{B_i(\theta_\gamma, \theta'_\gamma)}$ is representable as a function of the $I_{C_{ij}}$ indicators, so that

$$
\left| \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(B_i(\theta_\gamma, \theta'_\gamma)) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(B_i(\theta_\gamma, \theta'_\gamma)) \right|
$$

$$
= \left\| \mathbb{P}_{I_{B_i(\theta_\gamma, \theta'_\gamma)}(\mathcal{Z}_{tk}(\theta))} - \mathbb{P}_{I_{B_i(\theta_\gamma, \theta'_\gamma)}(\mathcal{Z}_{tk}(\theta'))} \right\|
$$

$$
\leq \left\| \mathbb{P}_{(I_{C_{i1}}(\mathcal{Z}_{tk}(\theta)),\ldots,I_{C_{i\ell_i(\theta_\gamma,\theta'_\gamma)}}(\mathcal{Z}_{tk}(\theta)))} - \mathbb{P}_{(I_{C_{i1}}(\mathcal{Z}_{tk}(\theta')),\ldots,I_{C_{i\ell_i(\theta_\gamma,\theta'_\gamma)}}(\mathcal{Z}_{tk}(\theta')))} \right\|
$$

$$
\leq 2^{\ell_i(\theta_\gamma,\theta'_\gamma)} \max_{J \subseteq \{1,\ldots,\ell_i(\theta_\gamma,\theta'_\gamma)\}} \mathbb{E}\left[\left(\prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_{tk}(\theta)) \right) \prod_{j \notin J} \left(1 - I_{C_{ij}}(\mathcal{Z}_{tk}(\theta)) \right) \right.
$$

$$
\left. - \left(\prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_{tk}(\theta')) \right) \prod_{j \notin J} \left(1 - I_{C_{ij}}(\mathcal{Z}_{tk}(\theta')) \right) \right]
$$

$$
\leq 2^{\ell_i(\theta_\gamma,\theta'_\gamma)} \sum_{J \subseteq \{1,\ldots,2^{\ell_i(\theta_\gamma,\theta'_\gamma)}\}} \left| \mathbb{E}\left[\prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_{tk}(\theta)) - \prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_{tk}(\theta')) \right] \right|
$$

$$
\leq 4^{\ell_i(\theta_\gamma,\theta'_\gamma)} \max_{J \subseteq \{1,\ldots,2^{\ell_i(\theta_\gamma,\theta'_\gamma)}\}} \left| \mathbb{E}\left[\prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_{tk}(\theta)) - \prod_{j \in J} I_{C_{ij}}(\mathcal{Z}_{tk}(\theta')) \right] \right|
$$

$$
= 4^{\ell_i(\theta_\gamma,\theta'_\gamma)} \max_{J \subseteq \{1,\ldots,2^{\ell_i(\theta_\gamma,\theta'_\gamma)}\}} \left| \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}\left(\bigcap_{j \in J} C_{ij} \right) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\left(\bigcap_{j \in J} C_{ij} \right) \right|.
$$

Note that $\bigcap_{j \in J} C_{ij}$ can be expressed as some $(A_1 \times (-\infty, t_1]) \times \cdots \times (A_k \times (-\infty, t_k])$, where each $A_p \in \mathcal{B}_{\mathcal{X}}$ and $t_p \in \mathbb{R}$, so that, letting $\hat{\ell} = \max_{\theta,\theta' \in \Theta_\gamma} \max_{i \leq \tilde{M}_k(\gamma)} \ell_i(\theta, \theta')$ and $\mathcal{C}_k = \{(A_1 \times (-\infty, t_1]) \times \cdots \times (A_k \times (-\infty, t_k]) : \forall j \leq k, A_j \in \mathcal{B}_{\mathcal{X}}, t_k \in \mathbb{R}\}$, this last expression is at most

$$
4^{\hat{\ell}} \sup_{C \in \mathcal{C}_k} \left| \mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(C) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(C) \right|.
$$

Next note that for any $C = (A_1 \times (-\infty, t_1]) \times \cdots \times (A_k \times (-\infty, t_k]) \in \mathcal{C}_k$, letting $C_1 = A_1 \times \cdots \times A_k$ and $C_2 = (-\infty, t_1] \times \cdots \times (-\infty, t_k]$,

$$
\mathbb{P}_{\mathcal{Z}_{tk}(\theta)}(C) - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}(C) = \mathbb{E}\left[\left(\mathbb{P}_{\mathbb{Y}_{tk}(\theta)|\mathbb{X}_{tk}}(C_2) - \mathbb{P}_{\mathbb{Y}_{tk}(\theta')|\mathbb{X}_{tk}}(C_2) \right) I_{C_1}(\mathbb{X}_{tk}) \right]
$$

$$
\leq \mathbb{E}\left[\left| \mathbb{P}_{\mathbb{Y}_{tk}(\theta)|\mathbb{X}_{tk}}(C_2) - \mathbb{P}_{\mathbb{Y}_{tk}(\theta')|\mathbb{X}_{tk}}(C_2) \right| \right].
$$

158

For $p \in \{1, \ldots, k\}$, let $C_{2p} = (-\infty, t_p]$. Then note that, by definition of d, for any given $x = (x_1, \ldots, x_k)$, the class $\mathcal{H}_x = \{x_p \mapsto I_{C_{2p}}(h(x_p)) : h \in \mathcal{F}\}$ is a VC class over $\{x_1, \ldots, x_k\}$ with VC dimension at most d. Furthremore, we have

$$\left| \mathbb{P}_{\mathbb{Y}_{tk}(\theta) | \mathbb{X}_{tk}}(C_2) - \mathbb{P}_{\mathbb{Y}_{tk}(\theta') | \mathbb{X}_{tk}}(C_2) \right|$$

$$= \left| \mathbb{P}_{(I_{C_{21}}(h^*_{t\theta}(X_{t1})), \ldots, I_{C_{2k}}(h^*_{t\theta}(X_{tk}))) | \mathbb{X}_{tk}} (\{(1, \ldots, 1)\}) \right.$$

$$\left. - \mathbb{P}_{(I_{C_{21}}(h^*_{t\theta'}(X_{t1})), \ldots, I_{C_{2k}}(h^*_{t\theta'}(X_{tk}))) | \mathbb{X}_{tk}} (\{(1, \ldots, 1)\}) \right|.$$

Therefore, the results of [Yang, Hanneke, and Carbonell, 2013] (in the proof of their Lemma 3) imply that

$$\left| \mathbb{P}_{\mathbb{Y}_{tk}(\theta) | \mathbb{X}_{tk}}(C_2) - \mathbb{P}_{\mathbb{Y}_{tk}(\theta') | \mathbb{X}_{tk}}(C_2) \right|$$

$$\leq 2^k \max_{y \in \{0,1\}^d} \max_{D \in \{1, \ldots, k\}^d} \left| \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta}(X_{tj}))\}_{j \in D} | \{X_{tj}\}_{j \in D}} (\{y\}) \right.$$

$$\left. - \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta'}(X_{tj}))\}_{j \in D} | \{X_{tj}\}_{j \in D}} (\{y\}) \right|.$$

Thus, we have

$$\mathbb{E} \left[\left| \mathbb{P}_{\mathbb{Y}_{tk}(\theta) | \mathbb{X}_{tk}}(C_2) - \mathbb{P}_{\mathbb{Y}_{tk}(\theta') | \mathbb{X}_{tk}}(C_2) \right| \right]$$

$$\leq 2^k \mathbb{E} \left[\max_{y \in \{0,1\}^d} \max_{D \in \{1, \ldots, k\}^d} \left| \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta}(X_{tj}))\}_{j \in D} | \{X_{tj}\}_{j \in D}} (\{y\}) \right. \right.$$

$$\left. \left. - \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta'}(X_{tj}))\}_{j \in D} | \{X_{tj}\}_{j \in D}} (\{y\}) \right| \right]$$

$$\leq 2^k \sum_{y \in \{0,1\}^d} \sum_{D \in \{1, \ldots, k\}^d} \mathbb{E} \left[\left| \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta}(X_{tj}))\}_{j \in D} | \{X_{tj}\}_{j \in D}} (\{y\}) \right. \right.$$

$$\left. \left. - \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta'}(X_{tj}))\}_{j \in D} | \{X_{tj}\}_{j \in D}} (\{y\}) \right| \right]$$

$$\leq 2^{d+k} k^d \max_{y \in \{0,1\}^d} \max_{D \in \{1, \ldots, k\}^d} \mathbb{E} \left[\left| \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta}(X_{tj}))\}_{j \in D} | \{X_{tj}\}_{j \in D}} (\{y\}) \right. \right.$$

$$\left. \left. - \mathbb{P}_{\{I_{C_{2j}}(h^*_{t\theta'}(X_{tj}))\}_{j \in D} | \{X_{tj}\}_{j \in D}} (\{y\}) \right| \right].$$

Exchangeability implies this is at most

$$2^{d+k}k^d \max_{y \in \{0,1\}^d} \sup_{t_1,\dots,t_d \in \mathbb{R}} \mathbb{E}\left[\left|\mathbb{P}_{\{I_{(-\infty,t_j]}(h_{t\theta}^*(X_{tj}))\}_{j=1}^d | \mathbb{X}_{td}}(\{y\})\right.\right.$$

$$\left.\left. - \mathbb{P}_{\{I_{(-\infty,t_j]}(h_{t\theta'}^*(X_{tj}))\}_{j=1}^d | \mathbb{X}_{td}}(\{y\})\right|\right]$$

$$= 2^{d+k}k^d \max_{y \in \{0,1\}^d} \sup_{t_1,\dots,t_d \in \mathbb{R}} \mathbb{E}\left[\left|\mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta))\}_{j=1}^d | \mathbb{X}_{td}}(\{y\})\right.\right.$$

$$\left.\left. - \mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta'))\}_{j=1}^d | \mathbb{X}_{td}}(\{y\})\right|\right].$$

[Yang, Hanneke, and Carbonell, 2013] argue that for all $y \in \{0,1\}^d$ and $t_1,\dots,t_d \in \mathbb{R}$,

$$\mathbb{E}\left[\left|\mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta))\}_{j=1}^d | \mathbb{X}_{td}}(\{y\}) - \mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta'))\}_{j=1}^d | \mathbb{X}_{td}}(\{y\})\right|\right]$$

$$\leq 4\sqrt{\left\|\mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta))\}_{j=1}^d,\mathbb{X}_{td}} - \mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta'))\}_{j=1}^d,\mathbb{X}_{td}}\right\|}.$$

Noting that

$$\left\|\mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta))\}_{j=1}^d,\mathbb{X}_{td}} - \mathbb{P}_{\{I_{(-\infty,t_j]}(Y_{tj}(\theta'))\}_{j=1}^d,\mathbb{X}_{td}}\right\| \leq \left\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\right\|$$

completes the proof. $\qquad\square$

We can use the above lemmas to design an estimator of $\pi_{\theta_\star}$. Specifically, we have the following result.

Theorem 9.4. *There exists an estimator* $\hat{\theta}_{T\theta_\star} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta_\star),\dots,\mathcal{Z}_{Td}(\theta_\star))$, *and functions* $R :$ $\mathbb{N}_0 \times (0,1] \to [0,\infty)$ *and* $\delta : \mathbb{N}_0 \times (0,1] \to [0,1]$ *such that, for any* $\alpha > 0$, $\lim_{T \to \infty} R(T,\alpha) =$ $\lim_{T \to \infty} \delta(T,\alpha) = 0$ *and for any* $T \in \mathbb{N}_0$ *and* $\theta_\star \in \Theta$,

$$\mathbb{P}\left(\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\| > R(T,\alpha)\right) \leq \delta(T,\alpha) \leq \alpha.$$

Proof. The estimator $\hat{\theta}_{T\theta_\star}$ we will use is precisely the minimum-distance skeleton estimate of $\mathbb{P}_{\mathcal{Z}_{td}(\theta_\star)}$ [Devroye and Lugosi, 2001, Yatracos, 1985]. [Yatracos, 1985] proved that if $N(\varepsilon)$ is

the ε-covering number of $\{\mathbb{P}_{\mathcal{Z}_{td}(\theta_\star)} : \theta \in \Theta\}$, then taking this $\hat{\theta}_{T\theta_\star}$ estimator, then for some $T_\varepsilon = O((1/\varepsilon^2) \log N(\varepsilon/4))$, any $T \geq T_\varepsilon$ has

$$\mathbb{E}\left[\left\|\mathbb{P}_{\mathcal{Z}_{td}(\hat{\theta}_{T\theta_\star})} - \mathbb{P}_{\mathcal{Z}_{td}(\theta_\star)}\right\|\right] < \varepsilon.$$

Thus, taking $G_T = \inf\{\varepsilon > 0 : T \geq T_\varepsilon\}$, we have

$$\mathbb{E}\left[\left\|\mathbb{P}_{\mathcal{Z}_{td}(\hat{\theta}_{T\theta_\star})} - \mathbb{P}_{\mathcal{Z}_{td}(\theta_\star)}\right\|\right] \leq G_T = o(1).$$

Letting $R'(T, \alpha)$ be any positive sequence with $G_T \ll R'(T, \alpha) \ll 1$ and $R'(T, \alpha) \geq G_T/\alpha$, and letting $\delta(T, \alpha) = G_T/R'(T, \alpha) = o(1)$, Markov's inequality implies

$$\mathbb{P}\left(\left\|\mathbb{P}_{\mathcal{Z}_{td}(\hat{\theta}_{T\theta_\star})} - \mathbb{P}_{\mathcal{Z}_{td}(\theta_\star)}\right\| > R'(T, \alpha)\right) \leq \delta(T, \alpha) \leq \alpha. \tag{9.1}$$

Letting $R(T, \alpha) = \min_k \left(M_k\left(R'(T, \alpha)\right) + r_k\right)$, since $R'(T, \alpha) = o(1)$ and $r_k = o(1)$, we have $R(T, \alpha) = o(1)$. Furthermore, composing (9.1) with Lemmas 9.1, 9.2, and 9.3, we have

$$\mathbb{P}\left(\left\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\right\| > R(T, \alpha)\right) \leq \delta(T, \alpha) \leq \alpha.$$

$\square$

Remark: Although the above result makes use of the minimum-distance skeleton estimator, which is typically not computationally efficient, it is often possible to achieve this same result (for certain families of distributions) using a simpler estimator, such as the maximum likelihood estimator. All we require is that the risk of the estimator converges to 0 at a known rate that is independent of $\theta_\star$. For instance, see [van de Geer, 2000b] for conditions on the family of distributions sufficient for this to be true of the maximum likelihood estimator.

9.3 Maximizing Customer Satisfaction in Combinatorial Auctions

We can use Theorem 9.4 in the context of various applications. For instance, consider the following application to the problem of serving a sequence of customers so as to maximize their

satisfaction.

Suppose there is a menu of n items $[n] = \{1, \ldots, n\}$, and each bundle $B \subseteq [n]$ has an associated price $p(B) \geq 0$. Suppose also there is a sequence of customers, each with a valuation function $v_t : 2^{[n]} \to \mathbb{R}$. We suppose these v_t functions are i.i.d. samples. We can then calculate the satisfaction function for each customer as $s_t(x)$, where $x \in \{0, 1\}^n$, and $s_t(x) = v_t(B_x) - p(B_x)$, where $B_x \subseteq [n]$ contains element $i \in [n]$ iff $x_i = 1$.

Now suppose we are able to ask each customer a number of questions before serving up a bundle $B_{\hat{x}_t}$ to that customer. More specifically, we are able to ask for the value $s_t(x)$ for any $x \in \{0, 1\}^n$. This is referred to as a *value query* in the literature on preference elicitation in combinatorial auctions (see Chapter 14 of [Cramton, Shoham, and Steinberg, 2006], [Zinkevich, Blum, and Sandholm, 2003]). We are interested in asking as few questions as possible, while satisfying the guarantee that $\mathbb{E}[s_t(\hat{x}_t) - \max_x s_t(x)] \leq \varepsilon$.

Now suppose, for every π and ε, we have a method $A(\pi, \varepsilon)$ such that, given that π is the actual distribution of the s_t functions, $A(\pi, \varepsilon)$ guarantees that the $\hat{x}_t$ value it selects has $\mathbb{E}[\max_x s_t(x) - s_t(\hat{x}_t)] \leq \varepsilon$; also let $\hat{N}_t(\pi, \varepsilon)$ denote the actual (random) number of queries the method $A(\pi, \varepsilon)$ would ask for the s_t function, and let $Q(\pi, \varepsilon) = \mathbb{E}[\hat{N}_t(\pi, \varepsilon)]$. We suppose the method never queries any $s_t(x)$ value twice for a given t, so that its number of queries for any given t is bounded.

Also suppose $\mathcal{F}$ is a VC subgraph class of functions mapping $\mathcal{X} = \{0, 1\}^n$ into $[-1, 1]$ with pseudodimension d, and that $\{\pi_\theta : \theta \in \Theta\}$ is a known totally bounded family of distributions over $\mathcal{F}$ such that the s_t functions have distribution $\pi_{\theta_\star}$ for some unknown $\theta_\star \in \Theta$. For any $\theta \in \Theta$ and $\gamma > 0$, let $\mathrm{B}(\theta, \gamma) = \{\theta' \in \Theta : \|\pi_\theta - \pi_{\theta'}\| \leq \gamma\}$.

Suppose, in addition to A, we have another method $A'(\varepsilon)$ that is not π-dependent, but still provides the ε-correctness guarantee, and makes a bounded number of queries (e.g., in the worst case, we could consider querying all 2^n points, but in most cases there are more clever π-independent methods that use far fewer queries, such as $O(1/\varepsilon^2)$). Consider the following

method; the quantities $\hat{\theta}_{T\theta_\star}$, $R(T, \alpha)$, and $\delta(T, \alpha)$ from Theorem 9.4 are here considered with respect P_X taken as the uniform distribution on $\{0, 1\}^n$.

Algorithm 2 An algorithm for sequentially maximizing expected customer satisfaction.
for $t = 1, 2, \ldots, T$ **do**

    Pick points $X_{t1}, X_{t2}, \ldots, X_{td}$ uniformly at random from $\{0, 1\}^n$

    if $R(t - 1, \varepsilon/2) > \varepsilon/8$ **then**

      Run $A'(\varepsilon)$

      Take $\hat{x}_t$ as the returned value

    else

      Let $\breve{\theta}_{t\theta_\star} \in \mathrm{B}\left(\hat{\theta}_{(t-1)\theta_\star}, R(t - 1, \varepsilon/2)\right)$ be such that

$$Q(\pi_{\breve{\theta}_{t\theta_\star}}, \varepsilon/4) \leq \min_{\theta \in \mathrm{B}\left(\hat{\theta}_{(t-1)\theta_\star}, R(t-1, \varepsilon/2)\right)} Q(\pi_\theta, \varepsilon/4) + 1/t$$

      Run $A(\pi_{\breve{\theta}_{t\theta_\star}}, \varepsilon/4)$ and let $\hat{x}_t$ be its return value

    end if

end for

The following theorem indicates that this method is correct, and furthermore that the long-run average number of queries is not much worse than that of a method that has direct knowledge of $\pi_{\theta_\star}$.

Theorem 9.5. *For the above method, $\forall t \leq T$, $\mathbb{E}[\max_x s_t(x) - s_t(\hat{x}_t)] \leq \varepsilon$. Furthermore, if $S_T(\varepsilon)$ is the total number of queries made by the method, then*

$$\limsup_{T \to \infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq Q(\pi_{\theta_\star}, \varepsilon/4) + d.$$

Proof. By Theorem 9.4, for any $t \leq T$, if $R(t - 1, \varepsilon/2) \leq \varepsilon/8$, then with probability at least $1 - \varepsilon/2$, $\|\pi_{\theta_\star} - \pi_{\hat{\theta}_{(t-1)\theta_\star}}\| \leq R(t - 1, \varepsilon/2)$, so that a triangle inequality implies $\|\pi_{\theta_\star} - \pi_{\breve{\theta}_{t\theta_\star}}\| \leq 2R(t - 1, \varepsilon/2) \leq \varepsilon/4$. Thus,

$$\mathbb{E}\left[\max_x s_t(x) - s_t(\hat{x}_t)\right]$$

$$\leq \mathbb{E}\left[\mathbb{E}\left[\max_x s_t(x) - s_t(\hat{x}_t)\Big|\breve{\theta}_{t\theta_\star}\right] \mathbb{1}\left[\|\pi_{\breve{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq \varepsilon/2\right]\right] + \varepsilon/2. \quad (9.2)$$

For $\theta \in \Theta$, let $\hat{x}_{t\theta}$ denote the point x that would be returned by $A(\pi_{\check{\theta}_{t\theta_\star}}, \varepsilon/4)$ when queries are answered by some $s_{t\theta} \sim \pi_\theta$ instead of s_t (and supposing $s_t = s_{t\theta_\star}$). If $\|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq \varepsilon/4$, then

$$\mathbb{E}\left[\max_x s_t(x) - s_t(\hat{x}_t) \Big| \check{\theta}_{t\theta_\star}\right] = \mathbb{E}\left[\max_x s_{t\theta_\star}(x) - s_{t\theta_\star}(\hat{x}_t) \Big| \check{\theta}_{t\theta_\star}\right]$$

$$\leq \mathbb{E}\left[\max_x s_{t\check{\theta}_{t\theta_\star}}(x) - s_{t\check{\theta}_{t\theta_\star}}(\hat{x}_{t\check{\theta}_{t\theta_\star}}) \Big| \check{\theta}_{t\theta_\star}\right] + \|\pi_{\check{\theta}_{t\theta_\star}} - \pi_{\theta_\star}\| \leq \varepsilon/4 + \varepsilon/4 = \varepsilon/2.$$

Plugging into (9.2), we have

$$\mathbb{E}\left[\max_x s_t(x) - s_t(\hat{x}_t)\right] \leq \varepsilon.$$

For the result on $S_T(\varepsilon)$, first note that $R(t-1, \varepsilon/2) > \varepsilon/8$ only finitely many times (due to $R(t, \alpha) = o(1)$), so that we can ignore those values of t in the asymptotic calculation (as the number of queries is always bounded), and rely on the correctness guarantee of A' for correctness. For the remaining t values, let N_t denote the number of queries made by $A(\pi_{\check{\theta}_{t\theta_\star}}, \varepsilon/4)$. then

$$\limsup_{T\to\infty} \frac{\mathbb{E}[S_T(\varepsilon)]}{T} \leq d + \limsup_{T\to\infty} \sum_{t=1}^{T} \mathbb{E}\left[N_t\right]/T.$$

Since

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[N_t \mathbb{1}[\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| > R(t-1, \varepsilon/2)]\right]$$

$$\leq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} 2^n \mathbb{P}\left(\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| > R(t-1, \varepsilon/2)\right)$$

$$\leq 2^n \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \delta(t-1, \varepsilon/2) = 0,$$

we have

$$\limsup_{T\to\infty} \sum_{t=1}^{T} \mathbb{E}\left[N_t\right]/T = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[N_t \mathbb{1}[\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| \leq R(t-1, \varepsilon/2)]\right].$$

For any $t \leq T$, let $N_t(\check{\theta}_{t\theta_\star})$ denote the number of queries $A(\pi_{\check{\theta}_{t\theta_\star}}, \varepsilon/4)$ would make if queries were answered with $s_{t\check{\theta}_{t\theta_\star}}$ instead of s_t. On the event $\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| \leq R(t-1, \varepsilon/2)$, we have

$$\mathbb{E}\left[N_t \Big| \check{\theta}_{t\theta_\star}\right] \leq \mathbb{E}\left[N_t(\check{\theta}_{t\theta_\star}) \Big| \check{\theta}_{t\theta_\star}\right] + 2R(t-1, \varepsilon/2)$$

$$= Q(\pi_{\check{\theta}_{t\theta_\star}}, \varepsilon/4) + 2R(t-1, \varepsilon/2) \leq Q(\pi_{\theta_\star}, \varepsilon/4) + 2R(t-1, \varepsilon/2) + 1/t.$$

164

Therefore,

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[N_t \mathbb{1}[\|\pi_{\hat{\theta}_{(t-1)\theta_\star}} - \pi_{\theta_\star}\| \le R(t-1, \varepsilon/2)] \right]$$

$$\le Q(\pi_{\theta_\star}, \varepsilon/4) + \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} 2R(t-1, \varepsilon/2) + 1/t = Q(\pi_{\theta_\star}, \varepsilon/4).$$

□

Note that in many cases, this result will even continue to hold with an infinite number of goods ($n = \infty$), since the general results of the previous section have no dependence on the cardinality of the space $\mathcal{X}$.

Chapter 10

Active Learning with a Drifting Distribution

Abstract

We study the problem of active learning in a stream-based setting, allowing the distribution of the examples to change over time. We prove upper bounds on the number of prediction mistakes and number of label requests for established disagreement-based active learning algorithms, both in the realizable case and under Tsybakov noise. We further prove minimax lower bounds for this problem.

10.1   Introduction

Most existing analyses of active learning are based on an i.i.d. assumption on the data. In this work, we assume the data are independent, but we allow the distribution from which the data are drawn to shift over time, while the target concept remains fixed. We consider this problem in a stream-based selective sampling model, and are interested in two quantities: the number of mistakes the algorithm makes on the first T examples in the stream, and the number of label

166

requests among the first T examples in the stream.

In particular, we study scenarios in which the distribution may drift within a fixed totally bounded family of distributions. Unlike previous models of distribution drift [Bartlett, 1992, Koby Crammer and Vaughan, 2010], the minimax number of mistakes (or excess number of mistakes, in the noisy case) can be *sublinear* in the number of samples.

We specifically study the classic CAL active learning strategy [Cohn, Atlas, and Ladner, 1994b] in this context, and bound the number of mistakes and label requests the algorithm makes in the realizable case, under conditions on the concept space and the family of possible distributions. We also exhibit lower bounds on these quantities that match our upper bounds in certain cases. We further study a noise-robust variant of CAL, and analyze its number of mistakes and number of label requests in noisy scenarios where the noise distribution remains fixed over time but the marginal distribution on $\mathcal{X}$ may shift. In particular, we upper bound these quantities under Tsybakov's noise conditions [Mammen and Tsybakov, 1999]. We also prove minimax lower bounds under these same conditions, though there is a gap between our upper and lower bounds.

10.2   Definition and Notations

As in the usual statistical learning problem, there is a standard Borel space $\mathcal{X}$, called the instance space, and a set $\mathbb{C}$ of measurable classifiers $h : \mathcal{X} \to \{-1, +1\}$, called the concept space. We additionally have a space $\mathbb{D}$ of distributions on $\mathcal{X}$, called the distribution space. Throughout, we suppose that the VC dimension of $\mathbb{C}$, denoted d below, is finite.

For any $\mu_1, \mu_2 \in \mathbb{D}$, let $\|\mu_1 - \mu_2\| = \sup_A \mu_1(A) - \mu_2(A)$ denote the total variation pseudo-distance between μ_1 and μ_2, where the set A in the $\sup$ ranges over all measurable subsets of $\mathcal{X}$. For any $\epsilon > 0$, let $\mathbb{D}_\epsilon$ denote a minimal ϵ-cover of $\mathbb{D}$, meaning that $\mathbb{D}_\epsilon \subseteq \mathbb{D}$ and $\forall \mu_1 \in \mathbb{D}, \exists \mu_2 \in \mathbb{D}_\epsilon$ s.t. $\|\mu_1 - \mu_2\| < \epsilon$, and that $\mathbb{D}_\epsilon$ has minimal possible size $|\mathbb{D}_\epsilon|$ among all subsets of $\mathbb{D}$ with this property.

In the learning problem, there is an unobservable sequence of distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots$, with

each $\mathcal{D}_t \in \mathbb{D}$, and an unobservable time-independent regular conditional distribution, which we represent by a function $\eta : \mathcal{X} \to [0, 1]$. Based on these quantities, we let $\mathcal{Z} = \{(X_t, Y_t)\}_{t=1}^{\infty}$ denote an infinite sequence of independent random variables, such that $\forall t, X_t \sim \mathcal{D}_t$, and the conditional distribution of Y_t given X_t satisfies $\forall x \in \mathcal{X}, \mathbb{P}(Y_t = +1 | X_t = x) = \eta(x)$. Thus, the joint distribution of (X_t, Y_t) is specified by the pair $(\mathcal{D}_t, \eta)$, and the distribution of $\mathcal{Z}$ is specified by the collection $\{\mathcal{D}_t\}_{t=1}^{\infty}$ along with η. We also denote by $\mathcal{Z}_t = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_t, Y_t)\}$ the first t such labeled examples. Note that the η conditional distribution is time-independent, since we are restricting ourselves to discussing drifting marginal distributions on $\mathcal{X}$, rather than drifting concepts. Concept drift is an important and interesting topic, but is beyond the scope of our present discussion.

In the active learning protocol, at each time t, the algorithm is presented with the value X_t, and is required to predict a label $\hat{Y}_t \in \{-1, +1\}$; then after making this prediction, it may optionally request to observe the true label value Y_t; as a means of book-keeping, if the algorithm requests a label Y_t on round t, we define $Q_t = 1$, and otherwise $Q_t = 0$.

We are primarily interested in two quantities. The first, $\hat{M}_T = \sum_{t=1}^{T} \mathbb{I}\left[\hat{Y}_t \neq Y_t\right]$, is the cumulative number of mistakes up to time T. The second quantity of interest, $\hat{Q}_T = \sum_{t=1}^{T} Q_t$, is the total number of labels requested up to time T. In particular, we will study the expectations of these quantities: $\bar{M}_T = \mathbb{E}\left[\hat{M}_T\right]$ and $\bar{Q}_T = \mathbb{E}\left[\hat{Q}_T\right]$. We are particularly interested in the asymptotic dependence of $\bar{Q}_T$ and $\bar{M}_T - \bar{M}_T^*$ on T, where $\bar{M}_T^* = \inf_{h \in \mathbb{C}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}[h(X_t) \neq Y_t]\right]$. We refer to $\bar{Q}_T$ as the expected number of label requests, and to $\bar{M}_T - \bar{M}_T^*$ as the expected excess number of mistakes. For any distribution P on $\mathcal{X}$, we define $\mathrm{er}_P(h) = \mathbb{E}_{X \sim P}[\eta(X)\mathbb{I}[h(X) = -1] + (1 - \eta(X))\mathbb{I}[h(X) = +1]]$, the probability of h making a mistake for $X \sim P$ and Y with conditional probability of being $+1$ equal $\eta(X)$. Note that, abbreviating $\mathrm{er}_t(h) = \mathrm{er}_{\mathcal{D}_t}(h) = \mathbb{P}(h(X_t) \neq Y_t)$, we have $\bar{M}_T^* = \inf_{h \in \mathbb{C}} \sum_{t=1}^{T} \mathrm{er}_t(h)$.

Scenarios in which both $\bar{M}_T - \bar{M}_T^*$ and $\bar{Q}_T$ are $o(T)$ (i.e., sublinear) are considered desirable, as these represent cases in which we do "learn" the proper way to predict labels, while asymp-

totically using far fewer labels than passive learning. Once establishing conditions under which this is possible, we may then further explore the trade-off between these two quantities.

We will additionally make use of the following notions. For $V \subseteq \mathbb{C}$, let $\mathrm{diam}_t(V) = \sup_{h,g \in V} \mathcal{D}_t(\{x : h(x) \neq g(x)\})$. For $h : \mathcal{X} \to \{-1,+1\}$, $\bar{\mathrm{er}}_{s:t}(h) = \frac{1}{t-s+1}\sum_{u=s}^{t} \mathrm{er}_u(h)$, and for finite $S \subseteq \mathcal{X} \times \{-1,+1\}$, $\hat{\mathrm{er}}(h;S) = \frac{1}{|S|}\sum_{(x,y)\in S} \mathbb{I}[h(x) \neq y]$. Also let $\mathbb{C}[S] = \{h \in \mathbb{C} : \hat{\mathrm{er}}(h;S) = 0\}$. Finally, for a distribution P on $\mathcal{X}$ and $r > 0$, define $\mathrm{B}_P(h,r) = \{g \in \mathbb{C} : P(x : h(x) \neq g(x)) \leq r\}$.

10.2.1 Assumptions

In addition to the assumption of independence of the X_t variables and that $d < \infty$, each result below is stated under various additional assumptions. The weakest such assumption is that $\mathbb{D}$ is *totally bounded*, in the following sense. For each $\epsilon > 0$, let $\mathbb{D}_\epsilon$ denote a minimal subset of $\mathbb{D}$ such that $\forall \mathcal{D} \in \mathbb{D}, \exists \mathcal{D}' \in \mathbb{D}_\epsilon$ s.t. $\|\mathcal{D} - \mathcal{D}'\| < \epsilon$: that is, a minimal ϵ-cover of $\mathbb{D}$. We say that $\mathbb{D}$ is totally bounded if it satisfies the following assumption.

Assumption 10.1. $\forall \epsilon > 0, |\mathbb{D}_\epsilon| < \infty$.

In some of the results below, we will be interested in deriving specific rates of convergence. Doing so requires us to make stronger assumptions about $\mathbb{D}$ than mere total boundedness. We will specifically consider the following condition, in which $c, m \in [0, \infty)$ are constants.

Assumption 10.2. $\forall \epsilon > 0, |\mathbb{D}_\epsilon| < c \cdot \epsilon^{-m}$.

For an example of a class $\mathbb{D}$ satisfying the total boundedness assumption, consider $\mathcal{X} = [0,1]^n$, and let $\mathbb{D}$ be the collection of distributions that have uniformly continuous density function with respect to the Lebesgue measure on $\mathcal{X}$, with modulus of continuity at most some value $\omega(\epsilon)$ for each value of $\epsilon > 0$, where $\omega(\epsilon)$ is a fixed real-valued function with $\lim_{\epsilon \to 0} \omega(\epsilon) = 0$.

As a more concrete example, when $\omega(\epsilon) = L\epsilon$ for some $L \in (0, \infty)$, this corresponds to the family of Lipschitz continuous density functions with Lipschitz constant at most L. In this case, we have $|\mathbb{D}_\epsilon| \leq O(\epsilon^{-n})$, satisfying Assumption 10.2.

10.3 Related Work

We discuss active learning under distribution drift, with fixed target concept. There are several branches of the literature that are highly relevant to this, including domain adaptation [Mansour, Mohri, and Rostamizadeh, 2008, 2009], online learning [Littlestone, 1988], learning with concept drift, and empirical processes for independent but not identically distributed data [van de Geer, 2000a].

Streamed-based Active Learning with a Fixed Distribution [Dasgupta, Kalai, and Monteleoni, 2009] show that a certain modified perceptron-like active learning algorithm can achieve a mistake bound $O(d \log(T))$ and query bound $\tilde{O}(d \log(T))$, when learning a linear separator under a uniform distribution on the unit sphere, in the realizable case. [Dekel, Gentile, and Sridharam, 2010] also analyze the problem of learning linear separators under a uniform distribution, but allowing Tsybakov noise. They find that with $\bar{Q}_T = \tilde{O}\left(d^{\frac{2\alpha}{\alpha+2}} T^{\frac{2}{\alpha+2}}\right)$ queries, it is possible to achieve an expected excess number of mistakes $\bar{M}_T - M_T^* = \tilde{O}\left(d^{\frac{\alpha+1}{\alpha+2}} \cdot T^{\frac{1}{\alpha+2}}\right)$. At this time, we know of no work studying the number of mistakes and queries achievable by active learning in a stream-based setting where the distribution may change over time.

Stream-based Passive Learning with a Drifting Distribution There has been work on learning with a drifting distribution and fixed target, in the context of passive learning. [Bartlett, 1992, Barve and Long, 1997] study the problem of learning a subset of a domain from randomly chosen examples when the probability distribution of the examples changes slowly but continually throughout the learning process; they give upper and lower bounds on the best achievable probability of misclassification after a given number of examples. They consider learning problems in which a changing environment is modeled by a slowly changing distribution on the product space. The allowable drift is restricted by ensuring that consecutive probability distributions are close in total variation distance. However, this assumption allows for certain malicious choices of distribution sequences, which shift the probability mass into smaller and smaller regions where

the algorithm is uncertain of the target's behavior, so that the number of mistakes grows linearly in the number of samples in the worst case. More recently, [Freund and Mansour, 1997] have investigated learning when the distribution changes as a linear function of time. They present algorithms that estimate the error of functions, using knowledge of this linear drift.

10.4 Active Learning in the Realizable Case

Throughout this section, suppose $\mathbb{C}$ is a fixed concept space and $h^* \in \mathbb{C}$ is a fixed target function: that is, $\mathrm{er}_t(h^*) = 0$. The family of scenarios in which this is true are often collectively referred to as the *realizable case*. We begin our analysis by studying this realizable case because it greatly simplifies the analysis, laying bare the core ideas in plain form. We will discuss more general scenarios, in which $\mathrm{er}_t(h^*) \geq 0$, in later sections, where we find that essentially the same principles apply there as in this initial realizable-case analysis.

We will be particularly interested in the performance of the following simple algorithm, due to [Cohn, Atlas, and Ladner, 1994b], typically referred to as CAL after its discoverers. The version presented here is specified in terms of a passive learning subroutine $\mathcal{A}$ (mapping any sequence of labeled examples to a classifier). In it, we use the notation $\mathrm{DIS}(V) = \{x \in \mathcal{X} : \exists h, g \in V \text{ s.t. } h(x) \neq g(x)\}$, also used below.

CAL

1. $t \leftarrow 0$, $\mathcal{Q}_0 \leftarrow \emptyset$, and let $\hat{h}_0 = \mathcal{A}(\emptyset)$

2. Do

3. $t \leftarrow t + 1$

4.   Predict $\hat{Y}_t = \hat{h}_{t-1}(X_t)$

5.   If $\max\limits_{y \in \{-1,+1\}} \min\limits_{h \in \mathbb{C}} \hat{\text{er}}(h; \mathcal{Q}_{t-1} \cup \{(X_t, y)\}) = 0$

6.   Request Y_t, let $\mathcal{Q}_t = \mathcal{Q}_{t-1} \cup \{(X_t, Y_t)\}$

7.   Else let $Y'_t = \operatorname*{argmin}\limits_{y \in \{-1,+1\}} \min\limits_{h \in \mathbb{C}} \hat{\text{er}}(h; \mathcal{Q}_{t-1} \cup \{(X_t, y)\})$, and let
$\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1} \cup \{(X_t, Y'_t)\}$

8.   Let $\hat{h}_t = \mathcal{A}(\mathcal{Q}_t)$

Below, we let $\mathcal{A}_{1IG}$ denote the one-inclusion graph prediction strategy of [Haussler, Littlestone, and Warmuth, 1994b]. Specifically, the passive learning algorithm $\mathcal{A}_{1IG}$ is specified as follows. For a sequence of data points $\mathcal{U} \in \mathcal{X}^{t+1}$, the one-inclusion graph is a graph, where each vertex represents a distinct labeling of $\mathcal{U}$ that can be realized by some classifier in $\mathbb{C}$, and two vertices are adjacent if and only if their corresponding labelings for $\mathcal{U}$ differ by exactly one label. We use the one-inclusion graph to define a classifier based on t training points as follows. Given t labeled data points $\mathcal{L} = \{(x_1, y_1), \ldots, (x_t, y_t)\}$, and one test point x_{t+1} we are asked to predict a label for, we first construct the one-inclusion graph on $\mathcal{U} = \{x_1, \ldots, x_{t+1}\}$; we then orient the graph (give each edge a unique direction) in a way that minimizes the maximum out-degree, and breaks ties in a way that is invariant to permutations of the order of points in $\mathcal{U}$; after orienting the graph in this way, we examine the subset of vertices whose corresponding labeling of $\mathcal{U}$ is consistent with $\mathcal{L}$; if there is only one such vertex, then we predict for x_{t+1} the corresponding label from that vertex; otherwise, if there are two such vertices, then they are adjacent in the one-inclusion graph, and we choose the one toward which the edge is directed and use the label for x_{t+1} in the corresponding labeling of $\mathcal{U}$ as our prediction for the label of x_{t+1}. See [Haussler, Littlestone, and Warmuth, 1994b] and subsequent work for detailed studies of the one-inclusion graph prediction strategy.

10.4.1 Learning with a Fixed Distribution

We begin the discussion with the simplest case: namely, when $|\mathbb{D}| = 1$.

Definition 10.3. *[Hanneke, 2007a, 2011] Define the disagreement coefficient of h^* under a distribution P as*

$$\theta_P(\epsilon) = \sup_{r > \epsilon} P\left(\mathrm{DIS}(\mathrm{B}_P(h^*, r))\right)/r.$$

Theorem 10.4. *For any distribution P on $\mathcal{X}$, if $\mathbb{D} = \{P\}$, then running CAL with $\mathcal{A} = \mathcal{A}_{1IG}$ achieves expected mistake bound $\bar{M}_T = O\left(d \log(T)\right)$ and expected query bound $\bar{Q}_T = O\left(\theta_P(\epsilon_T)d \log^2(T)\right)$, for $\epsilon_T = d \log(T)/T$.*

For completeness, the proof is included in the supplemental materials.

10.4.2 Learning with a Drifting Distribution

We now generalize the above results to any sequence of distributions from a totally bounded space $\mathbb{D}$. Throughout this section, let $\theta_{\mathbb{D}}(\epsilon) = \sup_{P \in \mathbb{D}} \theta_P(\epsilon)$.

First, we prove a basic result stating that CAL can achieve a sublinear number of mistakes, and under conditions on the disagreement coefficient, also a sublinear number of queries.

Theorem 10.5. *If $\mathbb{D}$ is totally bounded (Assumption 10.1), then CAL (with $\mathcal{A}$ any empirical risk minimization algorithm) achieves an expected mistake bound $\bar{M}_T = o(T)$, and if $\theta_{\mathbb{D}}(\epsilon) = o(1/\epsilon)$, then CAL makes an expected number of queries $\bar{Q}_T = o(T)$.*

Proof. As mentioned, given that $\mathrm{er}_{\mathcal{Q}_{t-1}}(h^*) = 0$, we have that Y'_t in Step 7 must equal $h^*(X_t)$, so that the invariant $\mathrm{er}_{\mathcal{Q}_t}(h^*) = 0$ is maintained for all t by induction. In particular, this implies $\mathcal{Q}_t = \mathcal{Z}_t$ for all t.

Fix any $\epsilon > 0$, and enumerate the elements of $\mathbb{D}_\epsilon$ so that $\mathbb{D}_\epsilon = \{P_1, P_2, \ldots, P_{|\mathbb{D}_\epsilon|}\}$. For each $t \in \mathbb{N}$, let $k(t) = \mathrm{argmin}_{k \leq |\mathbb{D}_\epsilon|} \|P_k - \mathcal{D}_t\|$, breaking ties arbitrarily. Let

$$L(\epsilon) = \left\lceil \frac{8}{\sqrt{\epsilon}}\left(d \ln\left(\frac{24}{\sqrt{\epsilon}}\right) + \ln\left(\frac{4}{\sqrt{\epsilon}}\right)\right)\right\rceil.$$

For each $i \le |\mathbb{D}_\epsilon|$, if $k(t) = i$ for infinitely many $t \in \mathbb{N}$, then let T_i denote the smallest value of T such that $|\{t \le T : k(t) = i\}| = L(\epsilon)$. If $k(t) = i$ only finitely many times, then let T_i denote the largest index t for which $k(t) = i$, or $T_i = 1$ if no such index t exists.

Let $T_\epsilon = \max_{i \le |\mathbb{D}_\epsilon|} T_i$ and $V_\epsilon = \mathbb{C}[\mathcal{Z}_{T_\epsilon}]$. We have that $\forall t > T_\epsilon, \mathrm{diam}_t(V_\epsilon) \le \mathrm{diam}_{k(t)}(V_\epsilon) + \epsilon$. For each i, let $\mathcal{L}_i$ be a sequence of $L(\epsilon)$ i.i.d. pairs (X, Y) with $X \sim P_i$ and $Y = h^*(X)$, and let $V_i = \mathbb{C}[\mathcal{L}_i]$. Then $\forall t > T_\epsilon$,

$$\mathbb{E}\left[\mathrm{diam}_{k(t)}(V_\epsilon)\right] \le \mathbb{E}\left[\mathrm{diam}_{k(t)}(V_{k(t)})\right] + \sum_{s \le T_i : k(s) = k(t)} \|\mathcal{D}_s - P_{k(s)}\| \le \mathbb{E}\left[\mathrm{diam}_{k(t)}(V_{k(t)})\right] + L(\epsilon)\epsilon.$$

By classic results in the theory of PAC learning [Anthony and Bartlett, 1999, Vapnik, 1982] and our choice of $L(\epsilon)$, $\forall t > T_\epsilon, \mathbb{E}\left[\mathrm{diam}_{k(t)}(V_{k(t)})\right] \le \sqrt{\epsilon}$.

Combining the above arguments,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] \le T_\epsilon + \sum_{t=T_\epsilon+1}^{T} \mathbb{E}\left[\mathrm{diam}_t(V_\epsilon)\right] \le T_\epsilon + \epsilon T + \sum_{t=T_\epsilon+1}^{T} \mathbb{E}\left[\mathrm{diam}_{k(t)}(V_\epsilon)\right]$$

$$\le T_\epsilon + \epsilon T + L(\epsilon)\epsilon T + \sum_{t=T_\epsilon+1}^{T} \mathbb{E}\left[\mathrm{diam}_{k(t)}(V_{k(t)})\right]$$

$$\le T_\epsilon + \epsilon T + L(\epsilon)\epsilon T + \sqrt{\epsilon}T.$$

Let ϵ_T be any nonincreasing sequence in $(0, 1)$ such that $1 \ll T_{\epsilon_T} \ll T$. Since $|\mathbb{D}_\epsilon| < \infty$ for all $\epsilon > 0$, we must have $\epsilon_T \to 0$. Thus, noting that $\lim_{\epsilon \to 0} L(\epsilon)\epsilon = 0$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] \le T_{\epsilon_T} + \epsilon_T T + L(\epsilon_T)\epsilon_T T + \sqrt{\epsilon_T}T \ll T. \tag{10.1}$$

The result on $\bar{M}_T$ now follows by noting that for any $\hat{h}_{t-1} \in \mathbb{C}[\mathcal{Z}_{t-1}]$ has $\mathrm{er}_t(\hat{h}_{t-1}) \le \mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])$, so

$$\bar{M}_T = \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{er}_t\left(\hat{h}_{t-1}\right)\right] \le \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] \ll T.$$

Similarly, for $r > 0$, we have

$$\mathbb{P}(\text{Request } Y_t) = \mathbb{E}\left[\mathbb{P}(X_t \in \mathrm{DIS}(\mathbb{C}[\mathcal{Z}_{t-1}])|\mathcal{Z}_{t-1})\right] \le \mathbb{E}\left[\mathbb{P}(X_t \in \mathrm{DIS}(\mathbb{C}[\mathcal{Z}_{t-1}] \cup \mathrm{B}_{\mathcal{D}_t}(h^*, r)))\right]$$

$$\le \mathbb{E}\left[\theta_{\mathbb{D}}(r) \cdot \max\left\{\mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}]), r\right\}\right] \le \theta_{\mathbb{D}}(r) \cdot r + \theta_{\mathbb{D}}(r) \cdot \mathbb{E}\left[\mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right].$$

174

Letting $r_T = T^{-1}\mathbb{E}\left[\sum_{t=1}^{T}\text{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right]$, we see that $r_T \to 0$ by (10.1), and since $\theta_{\mathbb{D}}(\epsilon) = o(1/\epsilon)$, we also have $\theta_{\mathbb{D}}(r_T)r_T \to 0$, so that $\theta_{\mathbb{D}}(r_T)r_T T \ll T$. Therefore, $\bar{Q}_T$ equals

$$\sum_{t=1}^{T}\mathbb{P}(\text{Request } Y_t) \leq \theta_{\mathbb{D}}(r_T)\cdot r_T\cdot T + \theta_{\mathbb{D}}(r_T)\cdot\mathbb{E}\left[\sum_{t=1}^{T}\text{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] = 2\theta_{\mathbb{D}}(r_T)\cdot r_T\cdot T \ll T. \quad \square$$

We can also state a more specific result in the case when we have some more detailed information on the sizes of the finite covers of $\mathbb{D}$.

Theorem 10.6. *If Assumption 10.2 is satisfied, then CAL (with $\mathcal{A}$ any empirical risk minimization algorithm) achieves an expected mistake bound $\bar{M}_T$ and expected number of queries $\bar{Q}_T$ such that $\bar{M}_T = O\left(T^{\frac{m}{m+1}}d^{\frac{1}{m+1}}\log^2 T\right)$ and $\bar{Q}_T = O\left(\theta_{\mathbb{D}}(\epsilon_T)T^{\frac{m}{m+1}}d^{\frac{1}{m+1}}\log^2 T\right)$, where $\epsilon_T = (d/T)^{\frac{1}{m+1}}$.*

Proof. Fix $\epsilon > 0$, enumerate $\mathbb{D}_\epsilon = \{P_1, P_2, \ldots, P_{|\mathbb{D}_\epsilon|}\}$, and for each $t \in \mathbb{N}$, define $k(t) = \text{argmin}_{1\leq k\leq|\mathbb{D}_\epsilon|}\|\mathcal{D}_t - P_k\|$. Let $\{X_t'\}_{t=1}^{\infty}$ be a sequence of independent samples, with $X_t' \sim P_{k(t)}$, and let $\mathcal{Z}_t' = \{(X_1', h^*(X_1')), \ldots, (X_t', h^*(X_t'))\}$. Then

$$\mathbb{E}\left[\sum_{t=1}^{T}\text{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\text{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}'])\right] + \sum_{t=1}^{T}\|\mathcal{D}_t - P_{k(t)}\|$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T}\text{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}'])\right] + \epsilon T \leq \sum_{t=1}^{T}\mathbb{E}\left[\text{diam}_{P_{k(t)}}(\mathbb{C}[\mathcal{Z}_{t-1}'])\right] + 2\epsilon T.$$

The classic convergence rates results from PAC learning [Anthony and Bartlett, 1999, Vapnik, 1982] imply

$$\sum_{t=1}^{T}\mathbb{E}\left[\text{diam}_{P_{k(t)}}(\mathbb{C}[\mathcal{Z}_{t-1}'])\right] = \sum_{t=1}^{T}O\left(\frac{d\log t}{|\{i\leq t:k(i)=k(t)\}|}\right)$$

$$\leq O(d\log T)\cdot\sum_{t=1}^{T}\frac{1}{|\{i\leq t:k(i)=k(t)\}|} \leq O(d\log T)\cdot|\mathbb{D}_\epsilon|\cdot\sum_{u=1}^{\lceil T/|\mathbb{D}_\epsilon|\rceil}\frac{1}{u} \leq O\left(d|\mathbb{D}_\epsilon|\log^2(T)\right).$$

Thus, $\sum_{t=1}^{T}\mathbb{E}\left[\text{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] \leq O\left(d|\mathbb{D}_\epsilon|\log^2(T) + \epsilon T\right) \leq O\left(d\cdot\epsilon^{-m}\log^2(T) + \epsilon T\right)$.
Taking $\epsilon = (T/d)^{-\frac{1}{m+1}}$, this is $O\left(d^{\frac{1}{m+1}}\cdot T^{\frac{m}{m+1}}\log^2(T)\right)$. We therefore have

$$\bar{M}_T \leq \mathbb{E}\left[\sum_{t=1}^{T}\sup_{h\in\mathbb{C}[\mathcal{Z}_{t-1}]}\text{er}_t(h)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\text{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] \leq O\left(d^{\frac{1}{m+1}}\cdot T^{\frac{m}{m+1}}\log^2(T)\right).$$

Similarly, letting $\epsilon_T = (d/T)^{\frac{1}{m+1}}$, $\bar{Q}_T$ is at most

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{D}_t(\mathrm{DIS}(\mathbb{C}[\mathcal{Z}_{t-1}]))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathcal{D}_t\left(\mathrm{DIS}\left(\mathrm{B}_{\mathcal{D}_t}\left(h^*, \max\left\{\mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}]), \epsilon_T\right\}\right)\right)\right)\right]$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \theta_{\mathbb{D}}\left(\epsilon_T\right) \cdot \max\left\{\mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}]), \epsilon_T\right\}\right]$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \theta_{\mathbb{D}}\left(\epsilon_T\right) \cdot \mathrm{diam}_t(\mathbb{C}[\mathcal{Z}_{t-1}])\right] + \theta_{\mathbb{D}}\left(\epsilon_T\right) T \epsilon_T \leq O\left(\theta_{\mathbb{D}}\left(\epsilon_T\right) \cdot d^{\frac{1}{m+1}} \cdot T^{\frac{m}{m+1}} \log^2(T)\right). \square$$

We can additionally construct a lower bound for this scenario, as follows. Suppose $\mathbb{C}$ contains a full infinite binary tree for which all classifiers in the tree agree on some point. That is, there is a set of points $\{x_b : b \in \{0,1\}^k, k \in \mathbb{N}\}$ such that, for $b_1 = 0$ and $\forall b_2, b_3, \ldots \in \{0,1\}$, $\exists h \in \mathbb{C}$ such that $h(x_{(b_1,\ldots,b_{j-1})}) = b_j$ for $j \geq 2$. For instance, this is the case for linear separators (and most other natural "geometric" concept spaces).

Theorem 10.7. *For any $\mathbb{C}$ as above, for any active learning algorithm, $\exists$ a set $\mathbb{D}$ satsifying Assumption 10.2, a target function $h^* \in \mathbb{C}$, and a sequence of distributions $\{\mathcal{D}_t\}_{t=1}^{T}$ in $\mathbb{D}$ such that the achieved $\bar{M}_T$ and $\bar{Q}_T$ satisfy $\bar{M}_T = \Omega\left(T^{\frac{m}{m+1}}\right)$, and $\bar{M}_T = O\left(T^{\frac{m}{m+1}}\right) \implies \bar{Q}_T = \Omega\left(T^{\frac{m}{m+1}}\right)$.*

The proof is analogous to that of Theorem 10.17 below, and is therefore omitted for brevity.

10.5   Learning with Noise

In this section, we extend the above analysis to allow for various types of noise conditions commonly studied in the literature. For this, we will need to study a noise-robust variant of CAL, below referred to as Agnostic CAL (or ACAL). We prove upper bounds achieved by ACAL, as well as (non-matching) minimax lower bounds.

10.5.1 Noise Conditions

The following assumption may be referred to as a *strictly benign noise* condition, which essentially says the model is specified correctly in that $h^* \in \mathbb{C}$, and though the labels may be stochastic, they are not completely random, but rather each is slightly biased toward the h^* label.

Assumption 10.8. $h^* = \mathrm{sign}(\eta - 1/2) \in \mathbb{C}$ *and* $\forall x, \eta(x) \neq 1/2$.

A particularly interesting special case of Assumption 10.8 is given by Tsybakov's noise conditions, which essentially control how common it is to have η values close to $1/2$. Formally:

Assumption 10.9. η *satisfies Assumption 10.8 and for some* $c > 0$ *and* $\alpha \geq 0$,
$$\forall t > 0, P(|\eta(x) - 1/2| < t) < c \cdot t^\alpha.$$

In the setting of shifting distributions, we will be interested in conditions for which the above assumptions are satisifed simultaneously for all distributions in $\mathbb{D}$. We formalize this in the following.

Assumption 10.10. *Assumption 10.9 is satisfied for all* $\mathcal{D} \in \mathbb{D}$, *with the same* c *and* α *values.*

10.5.2 Agnostic CAL

The following algorithm is essentially taken from [Dasgupta, Hsu, and Monteleoni, 2007a, Hanneke, 2011], adapted here for this stream-based setting. It is based on a subroutine: $\mathrm{LEARN}(\mathcal{L}, \mathcal{Q}) = \underset{h \in \mathbb{C} : \hat{\mathrm{er}}(h;\mathcal{L})=0}{\mathrm{argmin}} \hat{\mathrm{er}}(h;\mathcal{Q})$ if $\underset{h \in \mathbb{C}}{\min} \hat{\mathrm{er}}(h;\mathcal{L}) = 0$, and otherwise $\mathrm{LEARN}(\mathcal{L}, \mathcal{Q}) = \varnothing$.

ACAL

1. $t \leftarrow 0$, $\mathcal{L}_t \leftarrow \emptyset$, $\mathcal{Q}_t \leftarrow \emptyset$, let $\hat{h}_t$ be any element of $\mathbb{C}$

2. Do

3. $t \leftarrow t + 1$

4.   Predict $\hat{Y}_t = \hat{h}_{t-1}(X_t)$

5.   For each $y \in \{-1, +1\}$, let $h^{(y)} = \text{LEARN}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})$

6.   If either y has $h^{(-y)} = \varnothing$ or

    $\hat{\text{er}}(h^{(-y)}; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) - \hat{\text{er}}(h^{(y)}; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) > \hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})$

7. $\mathcal{L}_t \leftarrow \mathcal{L}_{t-1} \cup \{(X_t, y)\}$, $\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1}$

8.   Else Request Y_t, and let $\mathcal{L}_t \leftarrow \mathcal{L}_{t-1}$, $\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1} \cup \{(X_t, Y_t)\}$

9.   Let $\hat{h}_t = \text{LEARN}(\mathcal{L}_t, \mathcal{Q}_t)$

10.   If t is a power of 2

11. $\mathcal{L}_t \leftarrow \emptyset$, $\mathcal{Q}_t \leftarrow \emptyset$

The algorithm is expressed in terms of a function $\hat{\mathcal{E}}_t(\mathcal{L}, \mathcal{Q})$, defined as follows. Let δ_i be a nonincreasing sequence of values in $(0, 1)$. Let $\xi_1, \xi_2, \ldots$ denote a sequence of independent $Uniform(\{-1, +1\})$ random variables, also independent from the data. For $V \subseteq \mathbb{C}$, let $\hat{R}_t(V) = \sup_{h_1, h_2 \in V} \frac{1}{t - 2^{\lfloor \log_2(t-1) \rfloor}} \sum_{m=2^{\lfloor \log_2(t-1) \rfloor}+1}^{t} \xi_m \cdot (h_1(X_m) - h_2(X_m))$, $\hat{D}_t(V) = \sup_{h_1, h_2 \in V} \frac{1}{t - 2^{\lfloor \log_2(t-1) \rfloor}} \sum_{m=2^{\lfloor \log_2(t-1) \rfloor}+1}^{t} |h_1(X_m) - h_2(X_m)|$, $\hat{U}_t(V, \delta) = 12\hat{R}_t(V) + 34\sqrt{\hat{D}_t(V)\frac{\ln(32t^2/\delta)}{t}} + \frac{752\ln(32t^2/\delta)}{t}$. Also, for any finite sets $\mathcal{L}, \mathcal{Q} \subseteq \mathcal{X} \times \mathcal{Y}$, let $\mathbb{C}[\mathcal{L}] = \{h \in \mathbb{C} : \hat{\text{er}}(h; \mathcal{L}) = 0\}$, $\hat{\mathbb{C}}(\epsilon; \mathcal{L}, \mathcal{Q}) = \{h \in \mathbb{C}[\mathcal{L}] : \hat{\text{er}}(h; \mathcal{L} \cup \mathcal{Q}) - \min_{g \in \mathbb{C}[\mathcal{L}]} \hat{\text{er}}(g; \mathcal{L} \cup \mathcal{Q}) \leq \epsilon\}$. Then define $\hat{U}_t(\epsilon, \delta; \mathcal{L}, \mathcal{Q}) = \hat{U}_t(\hat{\mathbb{C}}_t(\epsilon; \mathcal{L}, \mathcal{Q}), \delta)$, and (letting $\mathbb{Z}_\epsilon = \{j \in \mathbb{Z} : 2^j \geq \epsilon\}$)

$$\hat{\mathcal{E}}_t(\mathcal{L}, \mathcal{Q}) = \inf\left\{\epsilon > 0 : \forall j \in \mathbb{Z}_\epsilon, \min_{m \in \mathbb{N}} \hat{U}_t(\epsilon, \delta_{\lfloor \log(t) \rfloor}; \mathcal{L}, \mathcal{Q}) \leq 2^{j-4}\right\}.$$

10.5.3 Learning with a Fixed Distribution

The following results essentially follow from [Hanneke, 2011], adapted to this stream-based setting.

Theorem 10.11. *For any strictly benign (P, η), if $2^{-2^i} \ll \delta_i \ll 2^{-i}/i$, ACAL achieves an expected excess number of mistakes $\bar{M}_T - M_T^* = o(T)$, and if $\theta_P(\epsilon) = o(1/\epsilon)$, then ACAL makes an expected number of queries $\bar{Q}_T = o(T)$.*

Theorem 10.12. *For any (P, η) satisfying Assumption 10.9, if $\mathbb{D} = \{P\}$, ACAL achieves an expected excess number of mistakes $\bar{M}_T - M_T^* = \tilde{O}\left(d^{\frac{1}{\alpha+2}} \cdot T^{\frac{\alpha+1}{\alpha+2}} \log\left(\frac{1}{\delta_{\lfloor \log(T) \rfloor}}\right) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i\right)$. and an expected number of queries $\bar{Q}_T = \tilde{O}\left(\theta_P(\epsilon_T) \cdot d^{\frac{2}{\alpha+2}} \cdot T^{\frac{\alpha}{\alpha+2}} \log\left(\frac{1}{\delta_{\lfloor \log(T) \rfloor}}\right) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i\right)$. where $\epsilon_T = T^{-\frac{\alpha}{\alpha+2}}$.*

Corollary 10.13. *For any (P, η) satisfying Assumption 10.9, if $\mathbb{D} = \{P\}$ and $\delta_i = 2^{-i}$ in ACAL, the algorithm achieves an expected number of mistakes $\bar{M}_T$ and expected number of queries $\bar{Q}_T$ such that, for $\epsilon_T = T^{-\frac{\alpha}{\alpha+2}}$, $\bar{M}_T - M_T^* = \tilde{O}\left(d^{\frac{1}{\alpha+2}} \cdot T^{\frac{\alpha+1}{\alpha+2}}\right)$, and $\bar{Q}_T = \tilde{O}\left(\theta_P(\epsilon_T) \cdot d^{\frac{2}{\alpha+2}} \cdot T^{\frac{\alpha}{\alpha+2}}\right)$.*

10.5.4 Learning with a Drifting Distribution

We can now state our results concerning ACAL, which are analogous to Theorems 10.5 and 10.6 proved earlier for CAL in the realizable case.

Theorem 10.14. *If $\mathbb{D}$ is totally bounded (Assumption 10.1) and η satisfies Assumption 10.8, then ACAL with $\delta_i = 2^{-i}$ achieves an excess expected mistake bound $\bar{M}_T - M_T^* = o(T)$, and if additionally $\theta_{\mathbb{D}}(\epsilon) = o(1/\epsilon)$, then ACAL makes an expected number of queries $\bar{Q}_T = o(T)$.*

The proof of Theorem 10.14 essentially follows from a combination of the reasoning for Theorem 10.5 and Theorem 10.15 below. Its proof is omitted.

Theorem 10.15. *If Assumptions 10.2 and 10.10 are satisfied, then ACAL achieves an expected excess number of mistakes $\bar{M}_T - M_T^* = \tilde{O}\left(T^{\frac{(\alpha+2)m+1}{(\alpha+2)(m+1)}} \log\left(\frac{1}{\delta_{\lfloor \log(T) \rfloor}}\right) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i\right)$, and an expected number of queries $\bar{Q}_T = \tilde{O}\left(\theta_{\mathbb{D}}(\epsilon_T) T^{\frac{(\alpha+2)(m+1)-\alpha}{(\alpha+2)(m+1)}} \log\left(\frac{1}{\delta_{\lfloor \log(T) \rfloor}}\right) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i\right)$, where $\epsilon_T = T^{-\frac{\alpha}{(\alpha+2)(m+1)}}$.*

The proof of this result is in many ways similar to that given above for the realizable case, and is included among the supplemental materials.

We immediately have the following corollary for a specific δ_i sequence.

Corollary 10.16. *With $\delta_i = 2^{-i}$ in ACAL, the algorithm achieves expected number of mistakes $\bar{M}$ and expected number of queries $\bar{Q}_T$ such that, for $\epsilon_T = T^{-\frac{\alpha}{(\alpha+2)(m+1)}}$,*

$$\bar{M}_T - M_T^* = \tilde{O}\left(T^{\frac{(\alpha+2)m+1}{(\alpha+2)(m+1)}}\right) \text{ and } \bar{Q}_T = \tilde{O}\left(\theta_{\mathbb{D}}(\epsilon_T) \cdot T^{\frac{(\alpha+2)(m+1)-\alpha}{(\alpha+2)(m+1)}}\right).$$

Just as in the realizable case, we can also state a minimax lower bound for this noisy setting.

Theorem 10.17. *For any $\mathbb{C}$ as in Theorem 10.7, for any active learning algorithm, $\exists$ a set $\mathbb{D}$ satisfying Assumption 10.2, a conditional distribution η, such that Assumption 10.10 is satisfied, and a sequence of distributions $\{\mathcal{D}_t\}_{t=1}^T$ in $\mathbb{D}$ such that the $\bar{M}_T$ and $\bar{Q}_T$ achieved by the learning algorithm satisfy $\bar{M}_T - M_T^* = \Omega\left(T^{\frac{1+m\alpha}{\alpha+2+m\alpha}}\right)$ and $\bar{M}_T - M_T^* = O\left(T^{\frac{1+m\alpha}{\alpha+2+m\alpha}}\right) \implies \bar{Q}_T = \Omega\left(T^{\frac{2+m\alpha}{\alpha+2+m\alpha}}\right).$*

The proof is included in the supplemental material.

10.6   Querying before Predicting

One interesting alternative to the above framework is to allow the learner to make a label request *before* making its label predictions. From a practical perspective, this may be more desirable and in many cases quite realistic. From a theoretical perspective, analysis of this alternative framework essentially separates out the mistakes due to over-confidence from the mistakes due to recognized uncertainty. In some sense, this is related to the KWIK model of learning of [Li, Littman, and Walsh, 2008].

Analyzing the above procedures in this alternative model yields several interesting details. Specifically, consider the following natural modifications to the above procedures. We refer to the algorithm LAC as the same sequence of steps as CAL, except with Step 4 removed, and

an additional step added after Step 8 as follows. In the case that we requested the label Y_t, we predict Y_t, and otherwise we predict $\hat{h}_t(X_t)$. Similarly, we define the algorithm ALAC as having the same sequence of steps as ACAL, except with Step 4 removed, and an additional step added after Step 11 as follows. In the case that we requested the label Y_t, we predict Y_t, and otherwise we predict $\hat{h}_t(X_t)$.

The analysis of the number of queries made by LAC in this setting remains essentially unchanged. However, if we consider running LAC in the realizable case, then the total number of mistakes in the entire sequence will be *zero*. As above, for any example for which LAC does not request the label, every classifier in the version space agrees with the target function's label, and therefore the inferred label will be correct. For any example that LAC requests the label of, in the setting where queries are made *before* predictions, we simply use the label itself as our prediction, so that LAC certainly does not make a mistake in this case.

On the other hand, the the analysis of ALAC in this alternative setting when we have noisy labels can be far more subtle. In particular, because the version space is only guaranteed to contain the best classifier *with high confidence*, there is still a small probability of making a prediction that disagrees with the best classifier h^* on each round that we do not request a label. So controlling the number of mistakes in this setting comes down to controlling the probability of removing h^* from the version space. However, this confidence parameter appears in the analysis of the number of queries, so that we have a natural trade-off between the number of mistakes and the number of label requests.

Formally, for any given nonincreasing sequence δ_i in $(0, 1)$, under Assumptions 10.2 and 10.10, ALAC achieves an expected excess number of mistakes $\bar{M}_T - M_T^* \leq \sum_{i=1}^{\lfloor \log(T) \rfloor} \delta_i 2^i$, and an expected number of queries $\bar{Q}_T = \tilde{O}\left(\theta_{\mathbb{D}}(\epsilon_T) \cdot T^{\frac{(\alpha+2)(m+1)-\alpha}{(\alpha+2)(m+1)}} \log\left(\frac{1}{\delta_{\lfloor \log(T) \rfloor}} \right) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i \right)$, where $\epsilon_T = T^{-\frac{\alpha}{(\alpha+2)(m+1)}}$. In particular, given any nondecreasing sequence M_T, we can set this δ_i sequence to maintain $\bar{M}_T - M_T^* \leq M_T$ for all T.

10.7 Discussion

What is not implied by the results above is any sort of *trade-off* between the number of mistakes and the number of queries. Intuitively, such a trade-off should exist; however, as CAL lacks any parameter to adjust the behavior with respect to this trade-off, it seems we need a different approach to address that question. In the batch setting, the analogous question is the trade-off between the number of label requests and the number of unlabeled examples needed. In the realizable case, that trade-off is tightly characterized by Dasgupta's *splitting index* analysis [Dasgupta, 2005]. It would be interesting to determine whether the splitting index tightly characterizes the mistakes-vs-queries trade-off in this stream-based setting as well.

In the batch setting, in which unlabeled examples are considered free, and performance is only measured as a function of the number of label requests, [Balcan, Hanneke, and Vaughan, 2010] have found that there is an important distinction between the *verifiable* label complexity and the *unverifiable* label complexity. In particular, while the former is sometimes no better than passive learning, the latter can always provide improvements for VC classes. Is there such a thing as unverifiable performance measures in the stream-based setting? To be concrete, we have the following open problem. Is there a method for every VC class that achieves $O(\log(T))$ mistakes and $o(T)$ queries in the realizable case?

10.8 Proof of Theorem 10.4

Proof of Theorem 10.4. First note that, by the assumption that $\forall t, \mathrm{er}_t(h^*) = 0$, with probability 1 we have that $\forall t, \mathcal{Q}_t = \mathcal{Z}_t$. Thus, since the stated bound on $\bar{M}_T$ for the one-inclusion graph algorithm has been established when using the true sequence of labeled examples $\mathcal{Z}_T$ [Haussler, Littlestone, and Warmuth, 1994b], it must hold here as well.

The remainder of the proof focuses on the bound on $\bar{Q}_T$. This proof is essentially based on a related proof of [Hanneke, 2011], but reformulated for this stream-based model.

Let V_t denote the set of classifiers $h \in \mathbb{C}$ with $\hat{\mathrm{er}}(h; \mathcal{Q}_t) = 0$ (with $V_0 = \mathbb{C}$). Classic results from statistical learning theory [Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989, Vapnik, 1982] imply that for $t > d$, with probability at least $1 - \delta$,

$$\mathrm{diam}_t(V_{t-1}) \leq cd \frac{\log(2e(t-1)/d) + \log(4/\delta)}{t-1}, \tag{10.2}$$

for some universal constant $c \in (1, \infty)$.

In particular, for $d < t \leq T$, since the probability CAL requests the label Y_t is $P(X_t \in \mathrm{DIS}(V_{t-1}))$, (10.2) implies that this probability satisfies

$$P\left(X_t \in \mathrm{DIS}(V_{t-1})\right) \leq P\left(X_t \in \mathrm{DIS}\left(\mathrm{B}_P\left(h^*, cd\frac{\log(2e(t-1)/d) + \log(4/\delta)}{t-1}\right)\right)\right) + \delta$$
$$\leq \theta_P\left(d\log(T)/T\right) cd\frac{\log(2e(t-1)/d) + \log(4/\delta)}{t-1} + \delta.$$

Taking $\delta = d/(t-1)$, this implies

$$P\left(X_t \in \mathrm{DIS}(V_{t-1})\right) \leq \theta_P\left(d\log(T)/T\right) 2cd\frac{\log(8e(t-1)/d)}{t-1}.$$

Thus, for $T > d$,

$$\bar{Q}_T = \sum_{t=1}^{T} P\left(X_t \in \mathrm{DIS}(V_{t-1})\right) \leq d + 1 + \sum_{t=d+1}^{T-1} \theta_P\left(d\log(T)/T\right) 2cd\frac{\log(8et/d)}{t}$$
$$\leq d + 1 + \theta_P\left(d\log(T)/T\right) 2cd\log(8eT/d) \int_d^T \frac{1}{t}\mathrm{d}t$$
$$= d + 1 + \theta_P\left(d\log(T)/T\right) 2cd\log(8eT/d)\log(T/d).$$

$\square$

10.9 Proof of Theorem 10.15

The following lemma is similar to a result proven by [Hanneke, 2011], based on the work of [Koltchinskii, 2006], except here we have adapted the result to the present setting with changing distributions. The proof is essentially identical to the proof of the original result of [Hanneke, 2011], and is therefore omitted here.

Lemma 10.18. *[Hanneke, 2011] Suppose η satisfies Assumption 10.8. For every $i \in \mathbb{N}$, on an event E_i with $\mathbb{P}(E_i) \geq 1 - \delta_i$, $\forall t \in \{2^i + 1, \ldots, 2^{i+1}\}$, letting $t(i) = t - 2^i$,*

- $\hat{\text{er}}(h^*; \mathcal{L}_{t-1}) = 0,$

- *$\forall h \in \mathbb{C}$ s.t. $\hat{\text{er}}(h; \mathcal{L}_{t-1}) = 0$ and $\hat{\text{er}}(h; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) - \hat{\text{er}}(h^*; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) \leq \hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1}),$*

   we have $\bar{\text{er}}_{2^i+1:t-1}(h) - \bar{\text{er}}_{2^i+1:t-1}(h^*) \leq 2\hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1}),$

- *if Assumption 10.10 is satisifed, $\hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1}) \leq \tilde{K} \cdot \left(\dfrac{d \log(t(i)/\delta_i)}{t(i)} \right)^{\frac{\alpha+1}{\alpha+2}},$*

for some (c, α)-dependent constant $\tilde{K} \in (1, \infty)$.

We can now prove Theorem 10.15.

Proof of Theorem 10.15. Fix any $i \in \mathbb{N}$, and we will focus on bounding the expected excess number of mistakes and expected number of queries for the values $t \in \{2^i + 1, \ldots, 2^{i+1}\}$. The result will then follow from this simply by summing this over values of $i \leq \log(T)$.

The predictions for $t \in \{2^i + 1, \ldots, 2^{i+1}\}$ are made by $\hat{h}_{t-1}$. Lemma 10.18 implies that with probability at least $1 - \delta_i$, every $t \in \{2^i + 1, \ldots, 2^{i+1}\}$ has $\forall h \in \mathbb{C}[\mathcal{L}_{t-1}]$ with $\hat{\text{er}}(h; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) - \hat{\text{er}}(h^*; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) \leq \hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})$ (and therefore in particular for $\hat{h}_{t-1}$)

$$\sum_{s=2^i+1}^{t-1} \text{er}_s(h) - \text{er}_s(h^*) \leq K_1 \cdot (t - 2^i) \cdot \left(\frac{d \log((t-2^i)/\delta_i)}{t - 2^i} \right)^{\frac{\alpha+1}{\alpha+2}}$$

$$\leq K_1 \cdot t^{\frac{1}{\alpha+2}} \cdot (d \log(t/\delta_i))^{\frac{\alpha+1}{\alpha+2}}. \qquad (10.3)$$

for some finite constant K_1.

Fix some value $\epsilon > 0$, and enumerate the elements of $\mathbb{D}_\epsilon = \{P_1, P_2, \ldots, P_{|\mathbb{D}_\epsilon|}\}$. Then let $\mathbb{D}_{\epsilon,k} = \{P \in \mathbb{D} : k = \text{argmin}_{j \leq |\mathbb{D}_\epsilon|} \|P_j - P\|\}$, breaking ties arbitrarily in the argmin. This induces a (Voronoi) partition $\{\mathbb{D}_{\epsilon,k} : k \leq |\mathbb{D}_\epsilon|\}$ of $\mathbb{D}$.

Rewriting (10.3) in terms of this partition, we have

$$\sum_{k=1}^{|\mathbb{D}_\epsilon|} \sum_{\substack{s \in \{2^i+1,\ldots,t-1\}: \\ \mathcal{D}_s \in \mathbb{D}_{\epsilon,k}}} \text{er}_s(h) - \text{er}_s(h^*) \leq K_1 \cdot (t)^{\frac{1}{\alpha+2}} \cdot (d \log(t/\delta_i)).$$

This means that, for any $k \leq |\mathbb{D}_\epsilon|$, we have

$$(\mathrm{er}_{P_k}(h) - \mathrm{er}_{P_k}(h^*)) \cdot \left|\left\{s \in \{2^i + 1, \ldots, t - 1\} : \mathcal{D}_s \in \mathbb{D}_{\epsilon,k}\right\}\right|$$

$$+ \sum_{s=2^i+1}^{t-1} (\mathrm{er}_s(h) - \mathrm{er}_s(h^*)) \cdot \mathbb{I}_{\mathbb{D}\backslash\mathbb{D}_{\epsilon,k}}(\mathcal{D}_s)$$

$$\leq K_1 \cdot (t)^{\frac{1}{\alpha+2}} \cdot (d\log(t/\delta_i)) + 2\epsilon \left|\left\{s \in \{2^i + 1, \ldots, t - 1\} : \mathcal{D}_s \in \mathbb{D}_{\epsilon,k}\right\}\right|.$$

Abbreviating by $k(s)$ the value of $k \leq |\mathbb{D}_\epsilon|$ with $\mathcal{D}_s \in \mathbb{D}_{\epsilon,k}$, we have that

$$\mathrm{er}_t(h) - \mathrm{er}_t(h^*)$$

$$\leq 2\epsilon + \mathrm{er}_{P_{k(t)}}(h) - \mathrm{er}_{P_{k(t)}}(h^*)$$

$$\leq 2\epsilon + \frac{2\epsilon \left|\{s \in \{2^i + 1, \ldots, t - 1\} : k(s) = k(t)\}\right| + K_1 \cdot (t)^{\frac{1}{\alpha+2}} \cdot (d\log(t/\delta_i))}{\max\left\{1, |\{s \in \{2^i + 1, \ldots, t - 1\} : k(s) = k(t)\}|\right\}}$$

$$\leq 4\epsilon + \frac{2K_1 \cdot (t)^{\frac{1}{\alpha+2}} \cdot (d\log(t/\delta_i))}{|\{s \in \{2^i + 1, \ldots, t\} : k(s) = k(t)\}|}. \tag{10.4}$$

Applying (10.4) simultaneously for all $t \in \{2^i + 1, \ldots, 2^{i+1}\}$ for $h = \hat{h}_{t-1}$, we have

$$\bar{M}_T - M_T^* \leq 4\epsilon T + \sum_{i=0}^{\lfloor \log(T) \rfloor} 2^i \delta_i +$$

$$2K_1 \cdot T^{\frac{1}{\alpha+2}} \cdot \log(T) \left(d\log(T/\delta_{\lfloor \log(T) \rfloor})\right) \sum_{i=0}^{\lfloor \log(T) \rfloor} \sum_{k=1}^{|\mathbb{D}_\epsilon|} \sum_{u=1}^{|\{t \in \{2^i+1,\ldots,2^{i+1}\}:k(t)=k\}|} \frac{1}{u}$$

$$\leq 4\epsilon T + \sum_{i=0}^{\lfloor \log(T) \rfloor} 2^i \delta_i +$$

$$2K_1 \cdot T^{\frac{1}{\alpha+2}} \cdot \log(T) \left(d\log(T/\delta_{\lfloor \log(T) \rfloor})\right) \log^2(2T) |\mathbb{D}_\epsilon|.$$

$$= O\left(\epsilon T + \epsilon^{-m} T^{\frac{1}{\alpha+2}} d\log^3(T) \log(1/\delta_{\lfloor \log(T) \rfloor}) + \sum_{i=0}^{\lfloor \log(T) \rfloor} 2^i \delta_i\right).$$

Taking $\epsilon = T^{-\frac{\alpha+1}{(\alpha+2)(m+1)}}$, this shows that

$$\bar{M}_T - M_T^* = O\left(T^{\frac{(\alpha+2)m+1}{(\alpha+2)(m+1)}} d\log^3(T) \log(1/\delta_{\lfloor \log(T) \rfloor}) + \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i\right).$$

185

We can bound $\bar{Q}_T$ in a similar fashion as follows. Fix any $i \leq \log(T)$. Lemma 10.18 implies that with probability at least $1 - \delta_i$, for every $t \in \{2^i + 1, \ldots, 2^{i+1}\}$, letting $\bar{\mathcal{E}}_t = 4\epsilon + \frac{2K_1 \cdot t^{\frac{1}{\alpha+2}} d \log(t/\delta_{\lfloor \log(t) \rfloor})}{|\{s \in \{2^i+1,\ldots,t\}: k(s)=k(t)\}|}$, we have

$$\mathbb{P}(\text{request } Y_t | \mathcal{L}_{t-1}, \mathcal{Q}_{t-1})$$

$$\leq \mathbb{P}\left(X_t \in \text{DIS}\left(\{h \in \mathbb{C}[\mathcal{L}_{t-1}] : \hat{\text{er}}(h; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) - \hat{\text{er}}(h^*; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) \leq \hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})\}\right) \Big| \mathcal{L}_{t-1}, \mathcal{Q}_{t-1}\right)$$

$$\leq \mathbb{P}\left(X_t \in \text{DIS}\left(\{h \in \mathbb{C} : \text{er}_t(h) - \text{er}_t(h^*) \leq \bar{\mathcal{E}}_t\}\right)\right)$$

$$\leq \mathbb{P}\left(X_t \in \text{DIS}\left(\left\{h \in \mathbb{C} : P_t(x : h(x) \neq h^*(x)) \leq K_2 \cdot \bar{\mathcal{E}}_t^{\frac{\alpha}{\alpha+1}}\right\}\right)\right)$$

$$\leq \theta_{\mathbb{D}}\left(\bar{\mathcal{E}}_t^{\frac{\alpha}{\alpha+1}}\right) \cdot K_3 \cdot \bar{\mathcal{E}}_t^{\frac{\alpha}{\alpha+1}},$$

where the third inequality above is due to Assumption 10.10.

Applying this simultaneously to all $i \leq \log(T)$ and $t \in \{2^i + 1, \ldots, 2^{i+1}\}$, we have, for $\bar{\epsilon}_T = \epsilon + T^{-\frac{\alpha+1}{\alpha+2}}$,

$$\bar{Q}_T \leq \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i + \theta_{\mathbb{D}}\left(\bar{\epsilon}_T^{\frac{\alpha}{\alpha+1}}\right) K_4 d \log(T/\delta_{\lfloor \log(T) \rfloor}) \sum_{i=0}^{\lfloor \log(T) \rfloor} \sum_{k=1}^{|\mathbb{D}_\epsilon|} \sum_{u=1}^{|\{t \in \{2^i+1,\ldots,2^{i+1}\}:k(t)=k\}|} \left(\max\left\{\epsilon, T^{\frac{1}{\alpha+2}}\frac{1}{u}\right\}\right)^{\frac{\alpha}{\alpha+1}}$$

$$\leq \sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i + \theta_{\mathbb{D}}\left(\bar{\epsilon}_T^{\frac{\alpha}{\alpha+1}}\right) \cdot K_5 \cdot d \log(1/\delta_{\lfloor \log(T) \rfloor}) \log^2(T) \cdot \left(\epsilon^{\frac{\alpha}{\alpha+1}} T + |\mathbb{D}_\epsilon| T^{\frac{\alpha}{(\alpha+2)(\alpha+1)}}\left(\frac{T}{|\mathbb{D}_\epsilon|}\right)^{\frac{1}{\alpha+1}}\right)$$

$$= O\left(\sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i + \theta_{\mathbb{D}}\left(\bar{\epsilon}_T^{\frac{\alpha}{\alpha+1}}\right) \log(1/\delta_{\lfloor \log(T) \rfloor}) \log^2(T) \cdot \left(\epsilon^{\frac{\alpha}{\alpha+1}} T + \epsilon^{-m\frac{\alpha}{\alpha+1}} T^{\frac{2}{\alpha+2}}\right)\right).$$

Taking $\epsilon = \epsilon_T^{\frac{\alpha+1}{\alpha}} = T^{-\frac{\alpha+1}{(\alpha+2)(m+1)}}$, we have

$$\bar{Q}_T = O\left(\sum_{i=0}^{\lfloor \log(T) \rfloor} \delta_i 2^i + \theta_{\mathbb{D}}(\epsilon_T) \log(1/\delta_{\lfloor \log(T) \rfloor}) \log^2(T) \cdot T^{\frac{(\alpha+2)(m+1)-\alpha}{(\alpha+2)(m+1)}}\right).$$

$\square$

10.10 Proof of Theorem 10.17

Proof of Theorem 10.17. Fix any $T \in \mathbb{N}$, and any particular active learning algorithm $\mathcal{A}$. We construct a set of distributions tailored for these, as follows. Let $\kappa = (\alpha + 1)/\alpha$. Let $\epsilon =$

186

$T^{-\frac{\kappa}{2\kappa-1+m}}$, $M = T^{\frac{m}{2\kappa+m-1}} = \epsilon^{-m/\kappa}$, and $K = T^{\frac{2\kappa-1}{2\kappa+m-1}} = T/M$.

Inductively define a sequence $\{b_k\}_{k=1}^{\infty}$ as follows. Let $b_1 = 0$, $b_2 = 1$. For any integer $k \geq 3$, given that values of $b_1, b_2, \ldots, b_{k-1}, \eta_3, \ldots, \eta_{k-1}, D_3, \ldots D_{k-1}$, and $X_1, X_2, \ldots, X_{(k-3)K}$ have already been defined, it is known [Hanneke, 2011] that for any active learning algorithm (possibly randomized) there exists a value b_k such that, for the distribution D_k with $D_k(\{x_{b_1,b_2,\ldots,b_{k-1}}\}) = \epsilon^{1/\kappa} = 1 - D_k(\{x_{b_1}\})$, there is a label distribution $\eta_k(x) = P(Y = 1|X = x)$ having $\eta_k(x_{b_1}) = 1$ and inducing $h^*(x_{b_1,b_2,\ldots,b_{k-1}}) = b_k$, which also satisfies Tsybakov noise with parameters c and α under distribution D_k: namely, $\eta_k(x_{b_1,b_2,\ldots,b_{k-1}}) = \frac{1}{2}\left(1 + (2b_k - 1)\epsilon^{\frac{\kappa-1}{\kappa}}\right)$. Furthermore, [Hanneke, 2011] shows that this b_k can be chosen so that, for some $N = \Omega\left(\epsilon^{\frac{2}{\kappa}-2}\right)$, after observing any number fewer than N random labeled observations (X, Y) with $X = x_{b_1,b_2,\ldots,b_{k-1}}$, if $\hat{h}_n$ is the algorithm's hypothesis, then $\mathbb{E}[\mathrm{er}(\hat{h}_n) - \mathrm{er}(h^*)] > \epsilon$, where the error rate is evaluated under η_k and D_k. In particular, this means that if the unlabeled samples are distributed according to D_k, then with any fewer than N label requests, the expected excess error rate will be greater than ϵ. But this also means that with any fewer than $\Omega(\epsilon^{-1/\kappa}N) = \Omega(\epsilon^{\frac{1}{\kappa}-2}) = \Omega(K)$ unlabeled examples sampled according to D_k, the expected excess error rate will be greater than ϵ.

Thus, to define the value b_k given the already-defined values $b_1, b_2, \ldots, b_{k-1}$, we consider $X_{(k-3)K+1}, X_{(k-3)K+2}, \ldots, X_{(k-2)K}$ i.i.d. D_k, independent from the other $X_1, \ldots, X_{(k-3)K}$ variables, and consider the values of b_k and η_k mentioned above, but defined for the active learning algorithm that feeds the stream $X_1, X_2, \ldots, X_{(k-3)K}$ into $\mathcal{A}$ before feeding in the samples from D_k. Thus, in this perspective, these $X_1, X_2, \ldots, X_{(k-3)K}$ random variables, and their labels (which $\mathcal{A}$ may request), are considered *internal* random variables in this active learning algorithm we have defined. This completes the inductive definition.

Now for the original learning problem we are interested in, we take as our fixed label distribution an η with $\eta(x_{b_1}) = 1$ and $\forall k \geq 2$, $\eta(x_{b_1,b_2,\ldots,b_{k-1}}) = \eta_k(x_{b_1,b_2,\ldots,b_{k-1}})$, and defined arbitrariliy elsewhere. Thus, for any D_k, this satisfies Tsybakov noise with the given c and α parameters.

We define the family $\mathbb{D}$ of distributions as $\{D_3, , D_4, \ldots, D_{M+2}\}$ for $M = T^{\frac{m}{2\kappa+m-1}} = \epsilon^{-m/\kappa}$

as above. Since these D_i are each separated by distance exactly $\epsilon^{1/\kappa}$, $\mathbb{D}$ satisfies the constraint on its cover sizes.

The sequence of data points will be the $X_1, X_2, \ldots, X_T$ sequence defined above, and the corresponding sequence of distributions has $\mathcal{D}_1 = \mathcal{D}_2 = \cdots = \mathcal{D}_K = D_3$, $\mathcal{D}_{K+1} = \mathcal{D}_{K+2} = \cdots = \mathcal{D}_{2K} = D_4$, and so on, up to $\mathcal{D}_{(M-1)K+1} = \mathcal{D}_{(M-1)K+2} = \cdots \mathcal{D}_T = D_{M+2}$.

Now applying the stated result of [Hanneke, 2011] used in the definition of the sequence, for any $1 \leq t \leq \min\{\epsilon^{-1/\kappa}N, K\}$, and any $k < M$, denoting by $\hat{h}_{kK+t-1}$ the classifier produced by $\mathcal{A}$ after processing $kK+t-1$ examples from this stream, $\mathbb{E}\left[\mathrm{er}_{\mathcal{D}_{kK+t}}(\hat{h}_{kK+t-1})\right] - \mathrm{er}_{\mathcal{D}_{kK+t}}(h^*) > \epsilon = T^{-\frac{\kappa}{2\kappa+m-1}}$.

Since $\min\{\epsilon^{-1/\kappa}N, K\} = \Omega(K)$, the expected excess number of mistakes is

$$
\begin{aligned}
\hat{M}_T - M_T^* &= \sum_{k=0}^{M-1}\sum_{t=1}^{K}\mathbb{E}\left[\mathrm{er}_{\mathcal{D}_{kK+t}}(\hat{h}_{kK+t-1})\right] - \mathrm{er}_{\mathcal{D}_{kK+t}}(h^*) \\
&\geq \sum_{k=0}^{M-1}\sum_{t=1}^{\min\{\epsilon^{-1/\kappa}N,K\}}\mathbb{E}\left[\mathrm{er}_{\mathcal{D}_{kK+t}}(\hat{h}_{kK+t-1})\right] - \mathrm{er}_{\mathcal{D}_{kK+t}}(h^*) \geq \sum_{k=0}^{M-1}\sum_{t=1}^{\min\{\epsilon^{-1/\kappa}N,K\}}\epsilon \\
&= \Omega\left(M \cdot K \cdot \epsilon\right) = \Omega\left(M \cdot (T/M) \cdot T^{-\frac{\kappa}{2\kappa+m-1}}\right) = \Omega\left(T^{\frac{\kappa+m-1}{2\kappa+m-1}}\right).
\end{aligned}
$$

Similarly, applying the stated result of [Hanneke, 2011] regarding the number of samples of labels for the point $x_{b_1,b_2,\ldots,b_{k-1}}$ to achieve excess error ϵ being larger than N, we see that in order to achieve this $\hat{M}_T - M_T^* = O\left(T^{\frac{\kappa+m-1}{2\kappa+m-1}}\right)$, we need that at least some constant fraction of these M segments receive an expected number of queries $\Omega(N)$, so that we will need $\hat{Q}_T = \Omega(M \cdot N) = \Omega\left(T^{\frac{2\kappa+m-2}{2\kappa+m-1}}\right)$. $\qquad\square$

Chapter 11

Active Learning with a Drifting Target Concept

Abstract

[1] This chapter describes results on learning in the presence of a drifting target concept. Specifically, we provide bounds on the expected number of mistakes on a sequence of i.i.d. points, labeled according to a target concept that can change by a given amount on each round. Some of the results also describe an active learning variant of this setting, and provide bounds on the number of queries for the labels of points in the sequence sufficient to obtain the stated bounds on the number of mistakes.

11.1 Introduction

At this time, the work on active learning has focused on learning settings in which the concept to be learned is static over time. However, in many real-world applications, such as webpage classification, spam filtering, and face recognition, the data distribution and the concept itself

---

[1]This chapter is based on joint work with Steve Hanneke and Varun Kanade.

change over time. Our existing work in the previous chapter addresses the problem of active learning with a drifting distribution, providing theoretical guarantees on the number of mistakes and label requests made by a particular active learning algorithm in a stream-based learning setting. However, that work left open the question of a drifting target concept. To bridge this gap, we propose to study the problem of active learning (and passive learning) with a drifting target concept. Specifically, consider a statistical learning setting, in which data arrive i.i.d. in a stream, and for each data point the learner is required to predict a label for the data point at that time, and then optionally request the true (target) label of that point. We are then interested in making a small number of queries and mistakes (including mistakes on unqueried labels) as a function of the number of points processed so far at any given time. The target labels are generated from a function known to reside in a given concept space, and at each time the target function is allowed to change by a distance ϵ (that is, the probability the new target function disagrees with the old target function on a random sample is at most ϵ). The recent work of [Koby Crammer and Vaughan, 2010] studies this problem in the context of passive learning of linear separators. In this theoretical study, we intend to broaden the scope of that work, to other concept spaces and distributions, improve the guarantees on performance, establish lower bounds on achievable performance, and extend the framework to study the number of labels requested by an active learning algorithm while maintaining the performance guarantees established for passive learning. In particular, we will be interested in bounding the number of queries and mistakes made by a particular algorithm, as a function of ϵ, the VC dimension of the concept space, and the number of time steps so far. We will also consider variants of this in which ϵ is also allowed to change over time, and then the bounds on the number of mistakes and queries should depend on the sequence of ϵ values.

11.2 Definitions and Notations

Formally, in this setting, there is a sequence of data i.i.d. unlabeled data $X_1, X_2, \ldots$, each with marginal distribution $\mathcal{P}$ over the instance space $\mathcal{X}$. There is also a sequence of target functions $h_1^*, h_2^*, \ldots$ in $\mathbb{C}$, with $\mathcal{P}(x : h_t^*(x) \neq h_{t+1}^*(x)) \leq \epsilon_{t+1}$ for each $t \in \mathbb{N}$. Each t has an associated target label $Y_t = h_t^*(X_t)$. A prediction $\hat{Y}_t$ is counted as a "mistake" if $\hat{Y}_t \neq Y_t$. We suppose each h_t^* is chosen independently from $X_t, X_{t+1}, \ldots$ (i.e., h_t^* is chosen prior to the "draw" of $X_t, X_{t+1}, \ldots \sim \mathcal{P}$). For the purposes of the results below, we do not necessarily require h_t^* to be independent from $X_1, \ldots, X_{t-1}$. Additionally, for any $x \in (0, \infty)$, define $\text{Log}(x) = \ln(x) \vee 1$.

11.3 General Analysis under Constant Drift Rate: Inefficient Passive Learning

The following Lemma is due to [Vapnik and Chervonenkis, 1971].

Lemma 11.1. *There exists a universal constant $c \in [1, \infty)$ such that, for any class $\mathbb{C}$ of VC dimension d, $\forall m \in \mathbb{N} \, \forall \delta \in (0, 1)$, with probability at least $1 - \delta$, every $h, g \in \mathbb{C}$ have*

$$
\left| \mathcal{P}(x : h(x) \neq g(x)) - \frac{1}{m} \sum_{t=1}^{m} \mathbb{I}[h(X_t) \neq g(X_t)] \right|
$$
$$
\leq c \sqrt{ \left(\frac{1}{m} \sum_{t=1}^{m} \mathbb{I}[h(X_t) \neq g(X_t)] \right) \frac{d \log(m/d) + \log(1/\delta)}{m} } + c \frac{d \log(m/d) + \log(1/\delta)}{m}.
$$

Consider the following algorithm.

> 0. Predict arbitrary values $\hat{Y}_1, \ldots, \hat{Y}_m$ for $Y_1, \ldots, Y_m$, respectively.
>
> 1. For $T = m + 1, m + 2, \ldots$
>
> 2.   Let $\hat{h}_T = \text{ERM}(\mathbb{C}, \{(X_{T-m}, Y_{T-m}), \ldots, (X_{T-1}, Y_{T-1})\})$
>
> 3.   Predict $\hat{Y}_T = \hat{h}_T(X_T)$ as the prediction for the value of Y_T

The bound in the following theorem is a generalization of one given by [Koby Crammer and Vaughan, 2010] for finite concept classes (which they claimed could be extended to spaces of

infinite VC dimension, presumably yielding something resembling the result stated here).

Theorem 11.2. *If every $\epsilon_t = \epsilon$, for some constant value $\epsilon \in (0,1)$, then the above algorithm, with $m = \lfloor \sqrt{d/\epsilon} \rfloor$, makes an expected number of mistakes among the first T instances that is $O(\sqrt{d\epsilon} \log(1/d\epsilon)T)$.*

Proof. The statement is trivial for any $\epsilon \geq 1/(ed)$, so suppose $\epsilon < 1/(ed)$. Let us bound $\mathrm{er}_t(\hat{h}_t) := \mathcal{P}(x : \hat{h}_t(x) \neq h_t^*(x))$ for an arbitrary $t > m$. By a Chernoff bound, with probability at least $1 - \delta$,

$$\frac{1}{m} \sum_{i=t-m}^{t-1} \mathbb{I}[h_{t-m}^*(X_i) \neq h_i^*(X_i)] \leq \frac{\log_2(1/\delta) + 2em^2\epsilon}{m} \leq (2\log_2(1/\delta) + 2ed)\sqrt{\epsilon/d}.$$

In particular, this means

$$\frac{1}{m} \sum_{i=t-m}^{t-1} \mathbb{I}[\hat{h}_t(X_i) \neq h_{t-m}^*(X_i)] \leq 2(2\log_2(1/\delta) + 2ed)\sqrt{\epsilon/d}.$$

By Lemma 11.1, on an additional event of probability at least $1 - \delta$,

$$\mathcal{P}(x : \hat{h}_t(x) \neq h_{t-m}^*(x))$$
$$\leq 2(2\log_2(1/\delta)+2ed)\sqrt{\epsilon/d}+c\sqrt{2(2\log_2(1/\delta) + 2ed)\sqrt{\epsilon/d}(d\log(1/\sqrt{d\epsilon}) + \log(1/\delta))2\sqrt{\epsilon/d}}$$
$$+ c(d\log(1/\sqrt{d\epsilon}) + \log(1/\delta))2\sqrt{\epsilon/d}.$$

Taking $\delta = \sqrt{d\epsilon}$, this is at most

$$2\sqrt{d\epsilon}\left((\sqrt{1/d}\log_2(1/d\epsilon) + 2e) + 2c\sqrt{1/d}\log_2(1/d\epsilon) + 2c\sqrt{2e\log(1/d\epsilon)} + c\log(1/d\epsilon)\right)$$
$$\leq 14(c+1)\sqrt{d\epsilon}\log(1/d\epsilon)$$

Since this holds with probability $1 - 2\delta = 1 - 2\sqrt{d\epsilon}$, and $\mathrm{er}_t(\hat{h}_t) \leq 1$ always, we have

$$\mathbb{E}\left[\mathrm{er}_t(\hat{h}_t)\right] \leq \mathcal{P}(x : \hat{h}_t(x) \neq h_{t-m}^*(x)) + \mathcal{P}(x : h_{t-m}^*(x) \neq h_t^*(x))$$
$$\leq 14(c+1)\sqrt{d\epsilon}\log(1/d\epsilon) + 2\sqrt{d\epsilon} + m\epsilon \leq (14c+17)\sqrt{d\epsilon}\log(1/d\epsilon).$$

Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}[\hat{h}_t(X_t) \neq h_t^*(X_t)]\right] \leq m + (14c+17)\sqrt{d\epsilon}\log(1/d\epsilon)(T-m) = O(\sqrt{d\epsilon}\log(1/d\epsilon)T).$$

$\square$

It may be possible to remove the $\log(1/d\epsilon)$ factor in some cases (e.g., homogeneous half-spaces under a uniform distribution on the sphere); it's not yet clear whether or not it should sometimes belong there in the optimal number of mistakes.

11.4 General Analysis under Constant Drift Rate: Sometimes-Efficient Passive Learning

The following method is often (though certainly not always) computationally efficient. For instance, it is efficient for linear separators.

---

0. Let $\hat{h}_0$ be an arbitrary classifier in $\mathbb{C}$

1. For $T = 1, 2, \ldots$

2.   If $T > m\lceil\log_2(1/\epsilon)\rceil$, let $m_T \in \{m, \ldots, m\lceil\log_2(1/\epsilon)\rceil\}$ be minimal s.t.

    $\min_{h\in\mathbb{C}} \sum_{t=T-m_T}^{T-m_T+m-1}\mathbb{I}[h(X_t) \neq Y_t] = 0$ (if it exists)

3.   If m_T exists, let $\hat{h}_T = \operatorname{argmin}_{h\in\mathbb{C}} \sum_{t=T-m_T}^{T-m_T+m-1}\mathbb{I}[h(X_t) \neq Y_t]$

4.   Else let $\hat{h}_T = \hat{h}_{T-1}$

5.   Predict $\hat{Y}_T = \hat{h}_T(X_T)$ as the prediction for the value of Y_T

Theorem 11.3. *If every $\epsilon_t = \epsilon$, for some constant value $\epsilon \in (0,1)$, then the above algorithm, with $m = \left\lfloor\frac{1}{2\sqrt{\epsilon}\lceil\log_2(1/\epsilon)\rceil}\right\rfloor$, makes an expected number of mistakes among the first T instances that is $O(d\sqrt{\epsilon}\log^2(1/\epsilon)T)$.*

Proof. The statement is trivial for any $\epsilon \geq 1/(ed)^2$, so suppose $\epsilon < 1/(ed)^2$. Let us bound $\mathbb{E}[\mathrm{er}_t(\hat{h}_t)] := \mathbb{E}[\mathcal{P}(x : \hat{h}_t(x) \neq h_t^*(x))]$ for an arbitrary $t > m\log_2(1/\sqrt{\epsilon})$.

Fix any $M \in \{m, \ldots, m\lceil \log_2(1/\epsilon)\rceil\}$. By a Chernoff bound, with probability at least $1 - \epsilon/(m\lceil \log_2(1/\epsilon)\rceil)$,

$$\frac{1}{m} \sum_{k=t-M}^{t-M+m-1} \mathbb{I}[h^*_{t-m\lceil \log_2(1/\epsilon)\rceil}(X_k) \neq h^*_k(X_k)] \leq \frac{1}{m} \log_2((m\lceil \log_2(1/\epsilon)\rceil)/\epsilon) + 2e\epsilon\lceil \log_2(1/\epsilon)\rceil m.$$

Combined with Lemma 11.1, this implies that with probability at least $1 - 2\epsilon/(m\lceil \log_2(1/\epsilon)\rceil)$, for any $h \in \mathbb{C}$ with

$$\sum_{k=t-M}^{t-M+m-1} \mathbb{I}[h(X_k) \neq h^*_k(X_k)] = 0,$$

it must have

$$\frac{1}{m} \sum_{k=t-M}^{t-M+m-1} \mathbb{I}[h^*_{t-m\lceil \log_2(1/\epsilon)\rceil}(X_k) \neq h(X_k)] \leq \frac{1}{m} \log_2((m\lceil \log_2(1/\epsilon)\rceil)/\epsilon) + 2e\epsilon\lceil \log_2(1/\epsilon)\rceil m,$$

and therefore

$$\mathcal{P}(x : h(x) \neq h^*_{t-m\lceil \log_2(1/\epsilon)\rceil}(x))$$

$$\leq \left(\frac{1}{m} \log_2((m\lceil \log_2(1/\epsilon)\rceil)/\epsilon) + 2e\epsilon\lceil \log_2(1/\epsilon)\rceil m \right)$$

$$+ c\sqrt{\left(\frac{1}{m} \log_2((m\lceil \log_2(1/\epsilon)\rceil)/\epsilon) + 2e\epsilon\lceil \log_2(1/\epsilon)\rceil m \right) \frac{d\log(m/d) + \log((m\lceil \log_2(1/\epsilon)\rceil)/\epsilon)}{m}}$$

$$+ c\frac{d\log(m/d) + \log((m\lceil \log_2(1/\epsilon)\rceil)/\epsilon)}{m}$$

$$\leq 19\sqrt{\epsilon}\log_2^2(1/\epsilon) + 12c\sqrt{d\epsilon}\log_2^2(1/\epsilon) + 24cd\sqrt{\epsilon}\log_2^2(1/\epsilon)$$

$$\leq 55cd\sqrt{\epsilon}\log_2^2(1/\epsilon).$$

If this is the case, then

$$\mathrm{er}_t(h) \leq \mathcal{P}(x : h^*_{t-m\lceil \log_2(1/\epsilon)\rceil}(x) \neq h_t(x)) + \mathcal{P}(x : h(x) \neq h^*_{t-m\lceil \log_2(1/\epsilon)\rceil}(x))$$

$$\leq \epsilon m\lceil \log_2(1/\epsilon)\rceil + 55cd\sqrt{\epsilon}\log_2^2(1/\epsilon)$$

$$\leq 56cd\sqrt{\epsilon}\log_2^2(1/\epsilon).$$

Thus, by a union bound, with probability at least $1 - 2\epsilon$, if m_t exists, then

$$\mathrm{er}_t(\hat{h}_t) \leq 56cd\sqrt{\epsilon}\log_2^2(1/\epsilon).$$

194

For any given $i \in \{1, \ldots, \lceil \log_2(1/\epsilon) \rceil\}$, by a union bound, the probability that

$$\sum_{k=t-mi}^{t-m(i-1)-1} \mathbb{I}[h^*_{t-m\lceil \log_2(1/\epsilon) \rceil}(X_k) \neq h^*_k(X_k)] > 0$$

is at most $\epsilon \lceil \log_2(1/\epsilon) \rceil m^2 < 1/2$. Since these sums are independent over values of i, we have

that with probability at least $1 - \epsilon$, at least one of these values of $i \in \{1, \ldots, \lceil \log_2(1/\epsilon) \rceil\}$ will

have $\sum_{k=t-mi}^{t-m(i-1)-1} \mathbb{I}[h^*_{t-m\lceil \log_2(1/\epsilon) \rceil}(X_k) \neq h^*_k(X_k)] = 0$. In particular, on this event, this implies

m_t exists in Step 2.

Altogether, since $\mathrm{er}_t(\hat{h}_t) \leq 1$ always, we have

$$\mathbb{E}[\mathrm{er}_t(\hat{h}_t)] \leq 56cd\sqrt{\epsilon}\log_2^2(1/\epsilon) + 3\epsilon \leq 59cd\sqrt{\epsilon}\log_2^2(1/\epsilon).$$

Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left[\hat{h}_t(X_t) \neq h^*_t(X_t)\right]\right] \leq m\lceil \log_2(1/\epsilon) \rceil + 59cd\sqrt{\epsilon}\log_2^2(1/\epsilon)T = O\left(d\sqrt{\epsilon}\log^2(1/\epsilon)T\right).$$

$\square$

11.4.1   Lower Bounds

In this section, we establish a lower bound on the number of mistakes that can be achieved when

the target function may drift by ϵ, at each step.

Thresholds

For simplicity, we first consider the case where the distribution is uniform over $[-1, 1]$, and the

concept class is threshold functions. Between each time-step the threshold may move to the left

or right by ϵ.

Theorem 11.4. *For any $\epsilon < 1/16$, any algorithm for learning under drifting targets makes at*

least $\sqrt{\epsilon}T/4e$ in expectation.

Proof. Consider the following strategy that the adversary uses to define the drifting thresholds. For simplicity assume that $2/\sqrt{\epsilon}$ is an even integer and T is divisible by $2/\sqrt{\epsilon}$. The game is divided into $k = T/(2/\sqrt{\epsilon})$ epochs, each consisting of $2/\sqrt{\epsilon}$ time steps. We have the following:

- At the beginning of each epoch, the threshold is at 0. The adverary tosses an unbiased coin.

- If the outcome is heads, for the next $1/\sqrt{\epsilon}$ time-steps, the threshold increase by ϵ at each time-step. Then for the next $1/\sqrt{\epsilon}$ it decreases by ϵ at each time-step. Thus, at the beginning of the next epoch, the threshold is again at 0.

- If the outcome is tails, the adversary first decreases the threshold by ϵ for the first $1/\sqrt{\epsilon}$ time-steps; then increases again. Thus, in either case, at the end of the epoch the threshold is again at 0.

We first assume that the algorithm knows the strategy of the adversary (but not the coin tosses). This can only make the algorithm more powerful. Since at the end of each epoch, the algorithm knows exactly where the threshold is, the total (expected) number of mistakes is k times the expected number of mistakes in each epoch. Without loss of generality consider the first epoch, *i.e.,* time-steps 1 to $2/\sqrt{\epsilon}$. For $t < \sqrt{t}$, let Z_t denote the random variable that is 1 if at time-step t, the random example x_t is inside the interval $[-\epsilon t, \epsilon t]$. Note that $\Pr[Z_t = 1] = \epsilon t$. Let M_t denote the random variable that is 1 if the algorithm makes a mistake at time-step t and 0 otherwise. (Here the expectation is over the randomness of the examples as well as the adversary's coin toss). Then, consider the following:

$$\mathbb{E}[M_t \mid Z_1 = 0, \ldots, Z_{t-1} = 0, Z_t = 1] = \frac{1}{2}$$

This is because, the only information the algorithm has at this time is that the threshold is either at $-\epsilon t$ or ϵt, each with equal probability. Therefore,

$$\mathbb{E}[M_t] \geq \frac{\epsilon t (1 - \sqrt{\epsilon})^{t-1}}{2}$$

196

Let $S = 1/\sqrt{\epsilon}$. Then, the expected number of mistakes between the time-steps 1 to S is $\mathbb{E}[\sum_{t=1}^{S} M_t] = \sum_{t=1}^{S} \mathbb{E}[M_t]$. Then, we have

$$\sum_{t=1}^{S} \mathbb{E}[M_t] \geq \frac{1}{2} \sum_{t=1}^{S} \epsilon t (1 - \sqrt{\epsilon})^{t-1}$$

Using the fact that $\sum_{t=1}^{S} t x^{t-1} \geq (1 - x^S)/(1 - x)$ for small enough x, we get

$$\sum_{t=1}^{S} \mathbb{E}[M_t] \geq \frac{\epsilon}{2} \cdot \frac{1 - (1 - \sqrt{\epsilon})^S}{(1 - (1 - \sqrt{\epsilon}))^2}$$

$$\geq \frac{1}{2e}$$

In the last line we used the fact that $(1 - x)^{1/x} \leq 1/e$. Now, it must be the case that the total (expected) number of mistakes is at least $k/2e = \sqrt{\epsilon}T/(4e)$. $\qquad \square$

Halfspaces

Now consider the case where $\mathcal{X} = \mathbb{R}^k$ for $k \in \mathbb{N}$, and where the concept space $\mathbb{C}$ is the set of halfspaces (linear separators): that is, for every $h \in \mathbb{C}$, $\exists w \in \mathbb{R}^k$ and $b \in \mathbb{R}$ such that $\forall x \in \mathbb{R}^k$, $h(x) = +1$ iff $w \cdot x + b \geq 0$. In this case, we have the following result.

Theorem 11.5. *For any $k \in \mathbb{N}$, for $\mathcal{X} = \mathbb{R}^k$ and $\mathbb{C}$ the class of halfspaces on $\mathbb{R}^k$, for any $\epsilon < 1/k$, for any algorithm for learning under ϵ-drifting targets, there exists a distribution $\mathcal{P}$ over $\mathbb{R}^k$ and a sequence of ϵ-drifting (w.r.t. $\mathcal{P}$) targets $h_1^*, h_2^*, \ldots$ in $\mathbb{C}$ such that, for any $T \in \mathbb{N}$, the expected number of mistakes made by the algorithm among the first T rounds is at least $\sqrt{\epsilon k}T/8$.*

Proof. Consider the distribution $\mathcal{P}$ that is uniform over the set

$$\bigcup_{i=1}^{k} \{0\}^{i-1} \times [0, 1] \times \{0\}^{k-i} :$$

that is, $\mathcal{P}$ is uniform in $[0, 1]$ along each of the axes. Now, by the probabilistic method, it suffices to show that there exists a way to randomly set the sequence of target functions so that the expected number of mistakes is at least the stated lower bound. We will choose the target functions

197

from among the subset of $\mathbb{C}$ consisting of halfspaces whose respective separating hyperplanes intersection all k axes in $[0, 1]$: that is, $\forall i \leq k$, $\{x : w \cdot x + b = 0\} \cap (\{0\}^{i-1} \times [0, 1] \times \{0\}^{k-i}) \neq \emptyset$. Note that each halfspace of this type can be specified by k values, $(z_1, \ldots, z_k)$, corresponding to the k intersection values with the axes: that is, $\forall i \leq k$, the $x \in \{0\}^{i-1} \times [0, 1] \times \{0\}^{k-i}$ has $x_i = z_i \in [0, 1]$.

Consider the following strategy that the adversary uses to define the drifting targets. For simplicity assume that $2\sqrt{k/\epsilon}$ is an even integer and T is divisible by $2\sqrt{k/\epsilon}$. The game is divided into $\ell = T/(2\sqrt{k/\epsilon})$ epochs, each consisting of $2\sqrt{k/\epsilon}$ time steps. We have the following:

- At the beginning of each epoch, the target function has $z_i = 1/2$ for all $i \leq k$. The adverary tosses k unbiased coins $c_1, \ldots, c_k$.

- For each $i \leq k$, if the outcome of tossing c_i is heads, for the next $\sqrt{k/\epsilon}$ time-steps, the value of z_i is increased by ϵ at each time-step, and then for the following $\sqrt{k/\epsilon}$ time-steps it decreases by ϵ. Thus, at the beginning of the next epoch, the target once again has $z_i = 1/2$ for all $i \leq k$.

- For each i, if the outcome of c_i is tails, the adversary first decreases the value of z_i by ϵ for the next $\sqrt{k/\epsilon}$ time-steps, and then increases again by ϵ on each round. Thus, in either case, at the end of the epoch the target again has $\forall i \leq k$, $z_i = 1/2$.

We first assume that the algorithm knows the strategy of the adversary (but not the coin tosses). This can only make the algorithm more powerful. Since at the end of each epoch, the algorithm knows exactly where the threshold is, the total (expected) number of mistakes is ℓ times the expected number of mistakes in each epoch. Without loss of generality consider the first epoch, *i.e.*, time-steps 1 to $2\sqrt{k/\epsilon}$. For $t \leq \sqrt{k/\epsilon}$ and $i \leq k$, let Z_{it} denote the random variable that is 1 if at time-step t, the ith coordinate of the random variable x_t is inside the interval $[1/2 - \epsilon t, 1/2 + \epsilon t]$. Note that $\Pr[Z_{it} = 1] = 2\epsilon t/k$. Let M_t denote the random variable that is 1 if the algorithm makes a mistake at time-step t and 0 otherwise. (Here the expectation is over the randomness of the examples as well as the adversary's coin tosses). Then, consider the

following:

$$\mathbb{E}[M_t \mid Z_{i1} = 0, \ldots, Z_{i(t-1)} = 0, Z_{it} = 1] = \frac{1}{2}.$$

For any $i \leq k$, if any $Z_{it} = 1$ for $t \leq \sqrt{k/\epsilon}$, then there must exist a first such t, in which case the above equality holds at that time t. Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{\sqrt{k/\epsilon}} M_t\right] \geq \sum_{i=1}^{k} \frac{1}{2} \mathbb{P}\left(\exists t \leq \sqrt{k/\epsilon} : Z_{it} = 1\right) = \frac{k}{2}\left(1 - \prod_{t=1}^{\sqrt{k/\epsilon}}(1 - 2\epsilon t/k)\right)$$

$$\geq \frac{k}{2}\left(1 - \exp\left\{-2(\epsilon/k)\sum_{t=1}^{\sqrt{k/\epsilon}} t\right\}\right) \geq \frac{k}{2}\left(1 - e^{-1}\right) \geq k/4.$$

Now, it must be the case that the total (expected) number of mistakes is at least $\ell k/4 = T\sqrt{\epsilon k}/8$.

$\square$

11.4.2 Random Drifts

In this section, we consider a very simple case of "random drift". We consider the class of homogeneous linear separators in $\mathbb{R}^2$, say $\mathbb{C}_2$ and let μ be any radially symmetric measure on $\mathbb{R}^2$.

We show a simple lower bound that the achievable target drift rate in this setting is $O(\epsilon^{2/3}T)$.

Proposition 11.6. *Let $\mathbb{C}_2$ be the class of homogeneous linear separators in $\mathbb{R}^2$ and let μ be any radially symmetric measure on $\mathbb{R}^2$. Then, if $c_1, c_2, \ldots, c_T$ is a (random) sequence of concepts from $\mathbb{C}_2$, where c_{i+1} is chosen uniformly at random from one of the two concepts in $\mathbb{C}_2$, such that $\mathrm{err}_\mu(c_i, c_{i+1}) = \epsilon$. Then, for any algorithm the expected number of mistakes is $\Omega(\epsilon^{2/3}T)$. (Here the expectation is taken over the randomness of the sequence c_i and the examples drawn from μ.)*

Proof. This follows from the anti-concentration of the standard random walk. $\square$

Proposition 11.7. *Under conditions of the above proposition – the algorithm above achieves a mistake bound of $O(\epsilon^{2/3}T)$.*

Proof. The main idea is that because of random drift, the expected number of examples that are consistent with a fixed classifier is actually $1/\epsilon^{1/3}$, instead of $1/\sqrt{\epsilon}$. □

11.5   Linear Separators under the Uniform Distribution

For the special case of learning linear separators in $\mathbb{R}^k$, the results of Section 11.4 imply that it is possible to achieve an expected number of mistakes and queries $\tilde{O}(d\sqrt{\epsilon}T)$ among the first T instances, using an algorithm that runs in time $\mathrm{poly}(d, 1/\epsilon)$ (and independent of T) for each prediction. In the special case of learning homogeneous linear separators under the uniform distribution on a unit sphere, it is possible to improve this result; specifically, we show there exists an efficient algorithm that achieves a bound on the expected number of mistakes and queries that is $\tilde{O}(\sqrt{d\epsilon}T)$, as was possible with the inefficient algorithm of Section 11.3. The technique is based on a modification of the algorithm presented in Section 11.3, replacing ERM with (a modification of) the computationally-efficient algorithm of [Awasthi, Balcan, and Long, 2013].

Formally, define the class of homogeneous linear separators as the set of classifiers $h_w :$ $\mathbb{R}^d \to \{-1, +1\}$, for $w \in \mathbb{R}^d$ with $\|w\| = 1$, such that $h_w(x) = \mathrm{sign}(w \cdot x)$ for every $x \in \mathbb{R}^d$. We have the following result.

Theorem 11.8. *When $\mathbb{C}$ is the space of homogeneous linear separators (with $d \geq 4$) and $\mathcal{P}$ is the uniform distribution on the surface of the origin-centered unit sphere in $\mathbb{R}^d$, when $\epsilon_t = \epsilon > 0$ (constant) for all $t \in \mathbb{N}$, there is an algorithm that runs in time $\mathrm{poly}(d, 1/\epsilon)$ for each prediction, which makes an expected number of mistakes among the first T instances that is $O\left(\sqrt{\epsilon d} \log^{3/2}\left(\frac{1}{\epsilon d}\right) T\right)$. Furthermore, the expected number of labels requested by the algorithm among the first T instances is $O\left(\sqrt{\epsilon d} \log^{3/2}\left(\frac{1}{\epsilon d}\right) T\right)$.*

Before stating the proof, we have a few additional definitions and lemmas that will be needed. For $\tau > 0$ and $x \in \mathbb{R}$, define $\ell_\tau(x) = \max\left\{0, 1 - \frac{x}{\tau}\right\}$. Consider the following algorithm and subroutine; parameters δ_k, m_k, τ_k, r_k, b_k, α, and κ will all be specified below; we suppose $M = \sum_{k=0}^{\lceil \log_2(1/\alpha) \rceil} m_k$.

Algorithm: DriftingHalfspaces
0. Let $\hat{w}_0$ be an arbitrary element of $\mathbb{R}^d$ with $\|\hat{w}_0\| = 1$
1. For $i = 1, 2, \ldots$
2.   ABL$(M(i-1))$

Subroutine: ModPerceptron(t)
0. Let w_t be any element of $\mathbb{R}^d$ with $\|w_t\| = 1$
1. For $m = t+1, t+2, \ldots, t+m_0$
2.   Predict $\hat{Y}_m = h_{w_{m-1}}(X_m)$ as the prediction for the value of Y_m
3.   Request the label Y_m
4.   If $\hat{Y}_m \neq Y_m$
5. $w_m \leftarrow w_{m-1} - 2(w_{m-1} \cdot X_m)X_m$
6.   Else $w_m \leftarrow w_{m-1}$
7. Return w_{t+m_0}

Subroutine: ABL(t)
0. Let w_0 be the return value of ModPerceptron(t)
1. For $k = 1, 2, \ldots, \lceil \log_2(1/\alpha) \rceil$
2. $W_k \leftarrow \{\}$
3.   For $s = t + \sum_{j=0}^{k-1} m_j + 1, \ldots, t + \sum_{j=0}^{k} m_j$
4.   Predict $\hat{Y}_s = h_{w_{k-1}}(X_s)$ as the prediction for the value of Y_s
5.   If $|w_{k-1} \cdot X_s| \leq b_{k-1}$, Request the label Y_s
6.   and let $W_k \leftarrow W_k \cup \{(X_s, Y_s)\}$
7.   Find $v_k \in \mathbb{R}^d$ with $\|v_k - w_{k-1}\| \leq r_k$, $0 < \|v_k\| \leq 1$, and
8. $\sum_{(x,y) \in W_k} \ell_{\tau_k}(y(v_k \cdot x)) \leq \inf_{v:\|v-w_{k-1}\| \leq r_k} \sum_{(x,y) \in W_k} \ell_{\tau_k}(y(v \cdot x)) + \kappa |W_k|$
9.   Let $w_k = \frac{1}{\|v_k\|} v_k$

The following result for ModPerceptron was proven by [Koby Crammer and Vaughan, 2010].

Lemma 11.9. *Suppose $\epsilon < \frac{1}{512}$. Consider the values w_m obtained during the execution of*

ModPerceptron(t). $\forall m \in \{t+1, \ldots, t+m_0\}$, $\mathcal{P}(x : h_{w_m}(x) \neq h_m^*(x)) \leq \mathcal{P}(x : h_{w_{m-1}}(x) \neq h_m^*(x))$. *Furthermore, letting* $c_1 = \frac{\pi^2}{d \cdot 400 \cdot 2^{15}}$, *if* $\mathcal{P}(x : h_{w_{m-1}}(x) \neq h_m^*(x)) \geq 1/32$, *then with probability at least* $1/64$, $\mathcal{P}(x : h_{w_m}(x) \neq h_m^*(x)) \leq (1-c_1)\mathcal{P}(x : h_{w_{m-1}}(x) \neq h_m^*(x))$.

This implies the following.

Lemma 11.10. *Suppose* $\epsilon \leq \frac{\pi^2}{400 \cdot 2^{27} d}$. *For* $m_0 = \max\left\{\left\lceil 512 \ln\left(\frac{1}{\sqrt{d\epsilon}}\right)\right\rceil, \lceil 128(1/c_1)\ln(32)\rceil\right\}$, *with probability at least* $1 - \sqrt{d\epsilon}$, *ModPerceptron(t) returns a vector* w *with* $\mathcal{P}(x : h_w(x) \neq h_{t+m_0+1}^*(x)) \leq 1/16$.

Proof. By Lemma 11.9 and a union bound, in general we have

$$\mathcal{P}(x : h_{w_m}(x) \neq h_{m+1}^*(x)) \leq \mathcal{P}(x : h_{w_{m-1}}(x) \neq h_m^*(x)) + \epsilon. \tag{11.1}$$

Furthermore, if $\mathcal{P}(x : h_{w_{m-1}}(x) \neq h_m^*(x)) \geq 1/32$, then wth probability at least $1/64$,

$$\mathcal{P}(x : h_{w_m}(x) \neq h_{m+1}^*(x)) \leq (1-c_1)\mathcal{P}(x : h_{w_{m-1}}(x) \neq h_m^*(x)) + \epsilon. \tag{11.2}$$

In particular, this implies that the number N of values $m \in \{t+1, \ldots, t+m_0\}$ with either $\mathcal{P}(x : h_{w_{m-1}}(x) \neq h_m^*(x)) < 1/32$ or $\mathcal{P}(x : h_{w_m}(x) \neq h_{m+1}^*(x)) \leq (1-c_1)\mathcal{P}(x : h_{w_{m-1}}(x) \neq h_m^*(x)) + \epsilon$ is lower-bounded by a $\text{Binomial}(m, 1/64)$ random variable. Thus, a Chernoff bound implies that with probability at least $1 - \exp\{-m_0/512\} \geq 1 - \sqrt{d\epsilon}$, we have $N \geq m_0/128$. Suppose this happens.

Since $\epsilon m_0 \leq 1/32$, if any $m \in \{t+1, \ldots, t+m_0\}$ has $\mathcal{P}(x : h_{w_{m-1}}(x) \neq h_m^*(x)) < 1/32$, then inductively applying (11.1) implies $\mathcal{P}(x : h_{w_{t+m_0}}(x) \neq h_{t+m_0+1}^*(x)) \leq 1/32 + \epsilon m_0 \leq 1/16$. On the other hand, if all $m \in \{t+1, \ldots, t+m_0\}$ have $\mathcal{P}(x : h_{w_{m-1}}(x) \neq h_m^*(x)) \geq 1/32$, then in particular we have N values of $m \in \{t+1, \ldots, t+m_0\}$ satisfying (11.2). Combining this fact with (11.1) inductively, we have that

$$\mathcal{P}(x : h_{w_{t+m_0}}(x) \neq h_{t+m_0+1}^*(x)) \leq (1-c_1)^N \mathcal{P}(x : h_{w_t}(x) \neq h_{t+1}^*(x)) + \epsilon m_0$$
$$\leq (1-c_1)^{(1/c_1)\ln(32)}\mathcal{P}(x : h_{w_t}(x) \neq h_{t+1}^*(x)) + \epsilon m_0 \leq \frac{1}{32} + \epsilon m_0 \leq \frac{1}{16}.$$

$\square$

Next, we consider the execution of $\mathrm{ABL}(t)$, and let the sets W_k be as in that execution. We will denote by w^* the weight vector with $\|w^*\| = 1$ such that $h^*_{t+m_0+1} = h_{w^*}$. Also denote by $M_1 = M - m_0$.

The proof relies on a few results proven in the work of [Awasthi, Balcan, and Long, 2013], which we summarize in the following lemmas. Although the results were proven in a slightly different setting in that work (namely, agnostic learning under a fixed joint distribution), one can easily verify that their proofs remain valid in our present context as well.

Lemma 11.11. *[Awasthi, Balcan, and Long, 2013] Fix any $k \in \{1, \ldots, \lceil \log_2(1/\alpha) \rceil\}$. Suppose $b_{k-1} = c_7 2^{1-k}/\sqrt{d}$ for a universal constant $c_7 > 0$, and let $z_k = \sqrt{r_k^2/(d-1) + b_{k-1}^2}$. For a universal constant $c_1 > 0$, if $\|w^* - w_{k-1}\| \leq r_k$,*

$$\left| \mathbb{E}\left[\sum_{(x,y)\in W_k} \ell_{\tau_k}(|w^* \cdot x|) \Big| w_{k-1}, |W_k| \right] - \mathbb{E}\left[\sum_{(x,y)\in W_k} \ell_{\tau_k}(y(w^* \cdot x)) \Big| w_{k-1}, |W_k| \right] \right|$$
$$\leq c_1 |W_k| \sqrt{2^k \epsilon M_1} \frac{z_k}{\tau_k}.$$

Lemma 11.12. *[Balcan and Long, 2013] For any $c > 0$, there is a constant $c' > 0$ depending only on c (i.e., not depending on d) such that, for any $u, v \in \mathbb{R}^d$ with $\|u\| = \|v\| = 1$, letting $\Delta = \mathcal{P}(x : h_u(x) \neq h_v(x))$, if $\Delta < 1/2$, then*

$$\mathcal{P}\left(x : h_u(x) \neq h_v(x) \text{ and } |v \cdot x| \geq c' \frac{\Delta}{\sqrt{d}} \right) \leq c\Delta.$$

The following is a well-known lemma concerning concentration around the equator for the uniform distribution (see e.g., [Awasthi, Balcan, and Long, 2013, Balcan, Broder, and Zhang, 2007b, Dasgupta, Kalai, and Monteleoni, 2009]); for instance, it easily follows from the formulas for the area in a spherical cap derived by [Li, 2011].

Lemma 11.13. *For any constant $C > 0$, there are constants $c_2, c_3 > 0$ depending only on C (i.e., independent of d) such that, for any $w \in \mathbb{R}^d$ with $\|w\| = 1$, $\forall \gamma \in [0, C/\sqrt{d}]$,*

$$c_2 \gamma \sqrt{d} \leq \mathcal{P}\left(x : |w \cdot x| \leq \gamma \right) \leq c_3 \gamma \sqrt{d}.$$

Based on this lemma, [Awasthi, Balcan, and Long, 2013] prove the following.

Lemma 11.14. *[Awasthi, Balcan, and Long, 2013] For $X \sim \mathcal{P}$, for any $w \in \mathbb{R}^d$ with $\|w\| = 1$, for any $C > 0$ and $\tau, b \in [0, C/\sqrt{d}]$, for c_2, c_3 as in Lemma 11.13,*

$$\mathbb{E}\left[\ell_\tau(|w^* \cdot X|)\Big| |w \cdot X| \le b\right] \le \frac{c_3 \tau}{c_2 b}.$$

The following is a slightly stronger version of a result of [Awasthi, Balcan, and Long, 2013] (specifically, the size of m_k, and consequently the bound on $|W_k|$, are both improved by a factor of d compared to the original result).

Lemma 11.15. *Fix any $\delta \in (0, 1/e)$. For universal constants $c_4, c_5, c_6, c_7, c_8, c_9, c_{10} \in (0, \infty)$, for an appropriate choice of $\kappa \in (0, 1)$ (a universal constant), if $\alpha = c_9\sqrt{\epsilon d \log\left(\frac{1}{\kappa\delta}\right)}$, for every $k \in \{1, \ldots, \lceil \log_2(1/\alpha) \rceil\}$, if $b_{k-1} = c_7 2^{1-k}/\sqrt{d}$, $\tau_k = c_8 2^{-k}/\sqrt{d}$, $r_k = c_{10} 2^{-k}$, $\delta_k = \delta/(\lceil \log_2(4/\alpha) \rceil - k)^2$, and $m_k = \left\lceil c_5 \frac{2^k}{\kappa^2} d \log\left(\frac{1}{\kappa\delta_k}\right)\right\rceil$, and if $\mathcal{P}(x : h_{w_{k-1}}(x) \ne h_{w^*}(x)) \le 2^{-k-3}$, then with probability at least $1 - (4/3)\delta_k$, $|W_k| \le c_6 \frac{1}{\kappa^2} d \log\left(\frac{1}{\kappa\delta_k}\right)$ and $\mathcal{P}(x : h_{w_k}(x) \ne h_{w^*}(x)) \le 2^{-k-4}$.*

Proof. By Lemma 11.13, and a Chernoff and union bound, for an appropriately large choice of c_5 and any $c_7 > 0$, letting c_2, c_3 be as in Lemma 11.13 (with $C = c_7 \vee (c_8/2)$), with probability at least $1 - \delta_k/3$,

$$c_2 c_7 2^{-k} m_k \le |W_k| \le 4 c_3 c_7 2^{-k} m_k. \tag{11.3}$$

The claimed upper bound on $|W_k|$ follows from this second inequality.

Next note that, if $\mathcal{P}(x : h_{w_{k-1}}(x) \ne h_{w^*}(x)) \le 2^{-k-3}$, then

$$\max\{\ell_{\tau_k}(y(w^* \cdot x)) : x \in \mathbb{R}^d, |w_{k-1} \cdot x| \le b_{k-1}, y \in \{-1, +1\}\} \le c_{11}\sqrt{d}$$

for some universal constant $c_{11} > 0$. Furthermore, since $\mathcal{P}(x : h_{w_{k-1}}(x) \ne h_{w^*}(x)) \le 2^{-k-3}$, we know that the angle between w_{k-1} and w^* is at most $2^{-k-3}\pi$, so that

$$\|w_{k-1} - w^*\| = \sqrt{2 - 2w_{k-1} \cdot w^*} \le \sqrt{2 - 2\cos(2^{-k-3}\pi)}$$

$$\le \sqrt{2 - 2\cos^2(2^{-k-3}\pi)} = \sqrt{2}\sin(2^{-k-3}\pi) \le 2^{-k-3}\pi\sqrt{2}.$$

For $c_{10} = \pi\sqrt{2}2^{-3}$, this is r_k. By Hoeffding's inequality (under the conditional distribution given $|W_k|$), the law of total probability, Lemma 11.11, and linearity of conditional expectations, with probability at least $1 - \delta_k/3$, for $X \sim \mathcal{P}$,

$$\sum_{(x,y)\in W_k} \ell_{\tau_k}(y(w^* \cdot x)) \leq |W_k|\mathbb{E}\left[\ell_{\tau_k}(|w^* \cdot X|)\Big| w_{k-1}, |w_{k-1} \cdot X| \leq b_{k-1}\right]$$

$$+ c_1|W_k|\sqrt{2^k\epsilon M_1}\frac{z_k}{\tau_k} + \sqrt{|W_k|(1/2)c_{11}^2 d\ln(3/\delta_k)}. \quad (11.4)$$

We bound each term on the right hand side separately. By Lemma 11.14, the first term is at most $|W_k|\frac{c_3\tau_k}{c_2 b_{k-1}} = |W_k|\frac{c_3 c_8}{2c_2 c_7}$. Next,

$$\frac{z_k}{\tau_k} = \frac{\sqrt{c_{10}^2 2^{-2k}/(d-1) + 4c_7^2 2^{-2k}/d}}{c_8 2^{-k}/\sqrt{d}} \leq \frac{\sqrt{2c_{10}^2 + 4c_7^2}}{c_8},$$

while $2^k \leq 2/\alpha$ so that the second term is at most

$$\sqrt{2}c_1 \frac{\sqrt{2c_{10}^2 + 4c_7^2}}{c_8}|W_k|\sqrt{\frac{\epsilon m}{\alpha}}.$$

Noting that

$$M_1 = \sum_{k'=1}^{\lceil\log_2(1/\alpha)\rceil} m_{k'} \leq \frac{32c_5}{\kappa^2}\frac{1}{\alpha}d\log\left(\frac{1}{\kappa\delta}\right), \quad (11.5)$$

we find that the second term on the right hand side of (11.4) is at most

$$\sqrt{\frac{c_5}{c_9}}\frac{8c_1}{\kappa}\frac{\sqrt{2c_{10}^2 + 4c_7^2}}{c_8}|W_k|\sqrt{\frac{\epsilon d\log\left(\frac{1}{\kappa\delta}\right)}{\alpha^2}} = \frac{8c_1\sqrt{c_5}}{\kappa}\frac{\sqrt{2c_{10}^2 + 4c_7^2}}{c_8 c_9}|W_k|.$$

Finally, since $d\ln(3/\delta_k) \leq 2d\ln(1/\delta_k) \leq \frac{2\kappa^2}{c_5}2^{-k}m_k$, and (11.3) implies $2^{-k}m_k \leq \frac{1}{c_2 c_7}|W_k|$, the third term on the right hand side of (11.4) is at most

$$|W_k|\frac{c_{11}\kappa}{\sqrt{c_2 c_5 c_7}}.$$

Altogether, we have

$$\sum_{(x,y)\in W_k} \ell_{\tau_k}(y(w^* \cdot x)) \leq |W_k|\left(\frac{c_3 c_8}{2c_2 c_7} + \frac{8c_1\sqrt{c_5}}{\kappa}\frac{\sqrt{2c_{10}^2 + 4c_7^2}}{c_8 c_9} + \frac{c_{11}\kappa}{\sqrt{c_2 c_5 c_7}}\right).$$

205

Taking $c_9 = 1/\kappa^3$ and $c_8 = \kappa$, this is at most

$$\kappa |W_k| \left(\frac{c_3}{2c_2 c_7} + 8c_1 \sqrt{c_5} \sqrt{2c_{10}^2 + 4c_7^2} + \frac{c_{11}}{\sqrt{c_2 c_5 c_7}} \right).$$

Next, note that because $h_{w_k}(x) \neq y \Rightarrow \ell_{\tau_k}(y(v_k \cdot x)) \geq 1$, and because (as proven above) $\|w^* - w_{k-1}\| \leq r_k$,

$$|W_k| \mathrm{er}_{W_k}(h_{w_k}) \leq \sum_{(x,y) \in W_k} \ell_{\tau_k}(y(v_k \cdot x)) \leq \sum_{(x,y) \in W_k} \ell_{\tau_k}(y(w^* \cdot x)) + \kappa |W_k|.$$

Combined with the above, we have

$$|W_k| \mathrm{er}_{W_k}(h_{w_k}) \leq \kappa |W_k| \left(1 + \frac{c_3}{2c_2 c_7} + 8c_1 \sqrt{c_5} \sqrt{2c_{10}^2 + 4c_7^2} + \frac{c_{11}}{\sqrt{c_2 c_5 c_7}} \right).$$

Let $c_{12} = 1 + \frac{c_3}{2c_2 c_7} + 8c_1 \sqrt{c_5} \sqrt{2c_{10}^2 + 4c_7^2} + \frac{c_{11}}{\sqrt{c_2 c_5 c_7}}$. Furthermore,

$$|W_k| \mathrm{er}_{W_k}(h_{w_k}) = \sum_{(x,y) \in W_k} \mathbb{I}[h_{w_k}(x) \neq y]$$

$$\geq \sum_{(x,y) \in W_k} \mathbb{I}[h_{w_k}(x) \neq h_{w^*}(x)] - \sum_{(x,y) \in W_k} \mathbb{I}[h_{w^*}(x) \neq y].$$

For an appropriately large value of c_5, by a Chernoff bound, with probability at least $1 - \delta_k/3$,

$$\sum_{s = t + \sum_{j=0}^{k-1} m_j + 1}^{t + \sum_{j=0}^{k} m_j} \mathbb{I}[h_{w^*}(X_s) \neq Y_s] \leq 2e\epsilon M_1 m_k + \log_2(3/\delta_k).$$

In particular, this implies

$$\sum_{(x,y) \in W_k} \mathbb{I}[h_{w^*}(x) \neq y] \leq 2e\epsilon M_1 m_k + \log_2(3/\delta_k),$$

so that

$$\sum_{(x,y) \in W_k} \mathbb{I}[h_{w_k}(x) \neq h_{w^*}(x)] \leq |W_k| \mathrm{er}_{W_k}(h_{w_k}) + 2e\epsilon M_1 m_k + \log_2(3/\delta_k).$$

Noting that (11.5) and (11.3) imply

$$\epsilon M_1 m_k \leq \epsilon \frac{32 c_5}{\kappa^2} \frac{d \log\left(\frac{1}{\kappa\delta}\right)}{c_9 \sqrt{\epsilon d \log\left(\frac{1}{\kappa\delta}\right)}} \frac{2^k}{c_2 c_7} |W_k| \leq \frac{32 c_5}{c_2 c_7 c_9 \kappa^2} \sqrt{\epsilon d \log\left(\frac{1}{\kappa\delta}\right)} 2^k |W_k|$$

$$= \frac{32 c_5}{c_2 c_7 c_9^2 \kappa^2} \alpha 2^k |W_k| = \frac{32 c_5 \kappa^4}{c_2 c_7} \alpha 2^k |W_k| \leq \frac{32 c_5 \kappa^4}{c_2 c_7} |W_k|,$$

and (11.3) implies $\log_2(3/\delta_k) \leq \frac{2\kappa^2}{c_2 c_5 c_7}|W_k|$, altogether we have

$$\sum_{(x,y)\in W_k} \mathbb{I}[h_{w_k}(x) \neq h_{w^*}(x)] \leq |W_k|\mathrm{er}_{W_k}(h_{w_k}) + \frac{64 e c_5 \kappa^4}{c_2 c_7}|W_k| + \frac{2\kappa^2}{c_2 c_5 c_7}|W_k|$$

$$\leq \kappa|W_k|\left(c_{12} + \frac{64 e c_5 \kappa^3}{c_2 c_7} + \frac{2\kappa}{c_2 c_5 c_7}\right).$$

Letting $c_{13} = c_{12} + \frac{64 e c_5}{c_2 c_7} + \frac{2}{c_2 c_5 c_7}$, and noting $\kappa \leq 1$, we have $\sum_{(x,y)\in W_k} \mathbb{I}[h_{w_k}(x) \neq h_{w^*}(x)] \leq c_{13}\kappa|W_k|$.

Lemma 11.1 (applied under the conditional distribution given $|W_k|$) and the law of total probability imply that with probability at least $1 - \delta_k/3$,

$$|W_k|\mathcal{P}\left(x : h_{w_k}(x) \neq h_{w^*}(x)\Big||w_{k-1} \cdot x| \leq b_{k-1}\right)$$

$$\leq \sum_{(x,y)\in W_k} \mathbb{I}[h_{w_k}(x) \neq h_{w^*}(x)] + c_{14}\sqrt{|W_k|(d\log(|W_k|/d) + \log(1/\delta_k))},$$

for a universal constant $c_{14} > 0$. Combined with the above, and the fact that (11.3) implies $\log(1/\delta_k) \leq \frac{\kappa^2}{c_2 c_5 c_7}|W_k|$ and

$$d\log(|W_k|/d) \leq d\log\left(\frac{8 c_3 c_5 c_7 \log\left(\frac{1}{\kappa\delta_k}\right)}{\kappa^2}\right)$$

$$\leq d\log\left(\frac{8 c_3 c_5 c_7}{\kappa^3 \delta_k}\right) \leq 3\log(8\max\{c_3, 1\}c_5)c_5 d\log\left(\frac{1}{\kappa\delta_k}\right)$$

$$\leq 3\log(8\max\{c_3, 1\})\kappa^2 2^{-k} m_k \leq \frac{3\log(8\max\{c_3, 1\})}{c_2 c_7}\kappa^2|W_k|,$$

we have

$$|W_k|\mathcal{P}\left(x : h_{w_k}(x) \neq h_{w^*}(x)\Big||w_{k-1} \cdot x| \leq b_{k-1}\right)$$

$$\leq c_{13}\kappa|W_k| + c_{14}\sqrt{|W_k|\left(\frac{3\log(8\max\{c_3, 1\})}{c_2 c_7}\kappa^2|W_k| + \frac{\kappa^2}{c_2 c_5 c_7}|W_k|\right)}$$

$$= \kappa|W_k|\left(c_{13} + c_{14}\sqrt{\frac{3\log(8\max\{c_3, 1\})}{c_2 c_7} + \frac{1}{c_2 c_5 c_7}}\right).$$

Thus, letting $c_{15} = \left(c_{13} + c_{14}\sqrt{\frac{3\log(8\max\{c_3,1\})}{c_2 c_7} + \frac{1}{c_2 c_5 c_7}}\right)$, we have

$$\mathcal{P}\left(x : h_{w_k}(x) \neq h_{w^*}(x)\Big||w_{k-1} \cdot x| \leq b_{k-1}\right) \leq c_{15}\kappa. \tag{11.6}$$

Next, note that $\|v_k - w_{k-1}\| = \sqrt{\|v_k\|^2 + 1 - 2\|v_k\| \cos(\pi \mathcal{P}(x : h_{w_k}(x) \neq h_{w_{k-1}}(x)))}$. Thus, one implication of the fact that $\|v_k - w_{k-1}\| \leq r_k$ is that $\frac{\|v_k\|}{2} + \frac{1 - r_k^2}{2\|v_k\|} \leq \cos(\pi \mathcal{P}(x : h_{w_k}(x) \neq h_{w_{k-1}}(x)))$; since the left hand side is positive, we have $\mathcal{P}(x : h_{w_k}(x) \neq h_{w_{k-1}}(x)) < 1/2$. Additionally, by differentiating, one can easily verify that for $\phi \in [0, \pi]$, $x \mapsto \sqrt{x^2 + 1 - 2x \cos(\phi)}$ is minimized at $x = \cos(\phi)$, in which case $\sqrt{x^2 + 1 - 2x \cos(\phi)} = \sin(\phi)$. Thus, $\|v_k - w_{k-1}\| \geq \sin(\pi \mathcal{P}(x : h_{w_k}(x) \neq h_{w_{k-1}}(x)))$. Since $\|v_k - w_{k-1}\| \leq r_k$, we have $\sin(\pi \mathcal{P}(x : h_{w_k}(x) \neq h_{w_{k-1}}(x))) \leq r_k$. Since $\sin(\pi x) \geq x$ for all $x \in [0, 1/2]$, combining this with the fact (proven above) that $\mathcal{P}(x : h_{w_k}(x) \neq h_{w_{k-1}}(x)) < 1/2$ implies $\mathcal{P}(x : h_{w_k}(x) \neq h_{w_{k-1}}(x)) \leq r_k$.

In particular, we have that both $\mathcal{P}(x : h_{w_k}(x) \neq h_{w_{k-1}}(x)) \leq r_k$ and $\mathcal{P}(x : h_{w^*}(x) \neq h_{w_{k-1}}(x)) \leq 2^{-k-3} \leq r_k$. Now Lemma 11.12 implies that, for any universal constant $c > 0$, there exists a corresponding universal constant $c' > 0$ such that

$$\mathcal{P}\left(x : h_{w_k}(x) \neq h_{w_{k-1}}(x) \text{ and } |w_{k-1} \cdot x| \geq c' \frac{r_k}{\sqrt{d}} \right) \leq c r_k$$

and

$$\mathcal{P}\left(x : h_{w^*}(x) \neq h_{w_{k-1}}(x) \text{ and } |w_{k-1} \cdot x| \geq c' \frac{r_k}{\sqrt{d}} \right) \leq c r_k,$$

so that (by a union bound)

$$\mathcal{P}\left(x : h_{w_k}(x) \neq h_{w^*}(x) \text{ and } |w_{k-1} \cdot x| \geq c' \frac{r_k}{\sqrt{d}} \right)$$
$$\leq \mathcal{P}\left(x : h_{w_k}(x) \neq h_{w_{k-1}}(x) \text{ and } |w_{k-1} \cdot x| \geq c' \frac{r_k}{\sqrt{d}} \right)$$
$$+ \mathcal{P}\left(x : h_{w^*}(x) \neq h_{w_{k-1}}(x) \text{ and } |w_{k-1} \cdot x| \geq c' \frac{r_k}{\sqrt{d}} \right)$$
$$\leq 2 c r_k.$$

In particular, letting $c_7 = c' c_{10}/2$, we have $c' \frac{r_k}{\sqrt{d}} = b_{k-1}$. Combining this with (11.6), Lemma 11.13,

and a union bound, we have that

$$\mathcal{P}\left(x : h_{w_k}(x) \neq h_{w^*}(x)\right)$$

$$\leq \mathcal{P}\left(x : h_{w_k}(x) \neq h_{w^*}(x) \text{ and } |w_{k-1} \cdot x| \geq b_{k-1}\right) + \mathcal{P}\left(x : h_{w_k}(x) \neq h_{w^*}(x) \text{ and } |w_{k-1} \cdot x| \leq b_{k-1}\right)$$

$$\leq 2cr_k + \mathcal{P}\left(x : h_{w_k}(x) \neq h_{w^*}(x) \Big| |w_{k-1} \cdot x| \leq b_{k-1}\right) \mathcal{P}\left(x : |w_{k-1} \cdot x| \leq b_{k-1}\right)$$

$$\leq 2cr_k + c_{15}\kappa c_3 b_{k-1}\sqrt{d} = \left(2^5 cc_{10} + c_{15}\kappa c_3 c_7 2^5\right) 2^{-k-4}.$$

Taking $c = \frac{1}{2^6 c_{10}}$ and $\kappa = \frac{1}{2^6 c_3 c_7 c_{15}}$, we have $\mathcal{P}(x : h_{w_k}(x) \neq h_{w^*}(x)) \leq 2^{-k-4}$, as required.

By a union bound, this occurs with probability at least $1 - (4/3)\delta_k$. $\qquad\square$

Proof of Theorem 11.8. If $\epsilon > \frac{\pi^2}{400 \cdot 2^{27} d}$, the result trivially holds, since then $T \leq \frac{400 \cdot 2^{27}}{\pi^2}\sqrt{\epsilon d}T$. Otherwise, suppose $\epsilon \leq \frac{\pi^2}{400 \cdot 2^{27} d}$.

Fix any $i \in \mathbb{N}$. Lemma 11.10 implies that, with probability at least $1 - \sqrt{\epsilon d}$, the w_0 returned in Step 0 of $\text{ABL}(M(i-1))$ satisfies $\mathcal{P}(x : h_{w_0}(x) \neq h^*_{M(i-1)+m_0+1}(x)) \leq 1/16$. Taking this as a base case, Lemma 11.15 (with $\delta = \sqrt{\epsilon d}$) then inductively implies that, with probability at least

$$1 - \sqrt{\epsilon d} - \sum_{k=1}^{\lceil \log_2(1/\alpha) \rceil} (4/3)\frac{\sqrt{\epsilon d}}{(\lceil \log_2(4/\alpha) \rceil - k)^2}$$

$$\geq 1 - \sqrt{\epsilon d}\left(1 + (4/3)\sum_{\ell=2}^{\infty}\frac{1}{\ell^2}\right) \geq 1 - 2\sqrt{\epsilon d},$$

every $k \in \{0, 1, \ldots, \lceil \log_2(1/\alpha) \rceil\}$ has

$$\mathcal{P}(x : h_{w_k}(x) \neq h^*_{M(i-1)+m_0+1}(x)) \leq 2^{-k-4}, \tag{11.7}$$

and furthermore the number of labels requested during $\text{ABL}(M(i-1))$ total to at most (for appropriate universal constants $\hat{c}_1, \hat{c}_2$)

$$m_0 + \sum_{k=1}^{\lceil \log_2(1/\alpha) \rceil} |W_k| \leq \hat{c}_1\left(d + \ln\left(\frac{1}{\epsilon d}\right) + \sum_{k=1}^{\lceil \log_2(1/\alpha) \rceil} d\log\left(\frac{(\lceil \log_2(4/\alpha) \rceil - k)^2}{\sqrt{\epsilon d}}\right)\right)$$

$$\leq \hat{c}_2 d\log^2\left(\frac{1}{\epsilon d}\right).$$

209

In particular, by a union bound, (11.7) implies that for every $k \in \{1, \ldots, \lceil \log_2(1/\alpha) \rceil\}$, every $m \in \left\{ M(i-1) + \sum_{j=0}^{k-1} m_j + 1, \ldots, M(i-1) + \sum_{j=0}^{k} m_j \right\}$ has

$$\mathcal{P}(x : h_{w_{k-1}}(x) \neq h_m^*(x))$$

$$\leq \mathcal{P}(x : h_{w_{k-1}}(x) \neq h_{M(i-1)+m_0+1}^*(x)) + \mathcal{P}(x : h_{M(i-1)+m_0+1}^*(x) \neq h_m^*(x))$$

$$\leq 2^{-k-3} + \epsilon M.$$

Thus, noting that

$$M = \sum_{k=0}^{\lceil \log_2(1/\alpha) \rceil} m_k = \Theta\left(d + \log\left(\frac{1}{\epsilon d}\right) + \sum_{k=1}^{\lceil \log_2(1/\alpha) \rceil} 2^k d \log\left(\frac{1}{\epsilon d}\right) \right)$$

$$= \Theta\left(\frac{1}{\alpha} d \log\left(\frac{1}{\epsilon d}\right) \right) = \Theta\left(\sqrt{\frac{d}{\epsilon}} \log\left(\frac{1}{\epsilon d}\right) \right),$$

we have that the expected number of labels requested among $\{y_{M(i-1)+1}, \ldots, y_{Mi}\}$ is at most

$$\hat{c}_2 d \log^2\left(\frac{1}{\epsilon d}\right) + 2\sqrt{\epsilon d} M = O\left(\sqrt{\epsilon d} \log^{3/2}\left(\frac{1}{\epsilon d}\right) M \right),$$

and the expected number of mistaken predictions among points $\{x_{M(i-1)+1}, \ldots, x_{Mi}\}$ is at most

$$2\sqrt{\epsilon d} M + (1 - 2\sqrt{\epsilon d}) \left(m_0 + \sum_{k=1}^{\lceil \log_2(1/\alpha) \rceil} (2^{-k-3} + \epsilon M) m_k \right)$$

$$= O\left(\sqrt{\epsilon d} M + d \log^2\left(\frac{1}{\epsilon d}\right) + \epsilon M^2 \right) = O\left(\sqrt{\epsilon d} \log^{3/2}\left(\frac{1}{\epsilon d}\right) M \right).$$

These imply that the expected number of labels requested among $\{y_1, \ldots, y_T\}$, for any given T, is at most

$$O\left(\sqrt{\epsilon d} \log^{3/2}\left(\frac{1}{\epsilon d}\right) M \left\lceil \frac{T}{M} \right\rceil \right) = O\left(\sqrt{\epsilon d} \log^{3/2}\left(\frac{1}{\epsilon d}\right) T \right),$$

and the expected number of mistaken predictions among points $\{x_1, \ldots, x_T\}$ is at most

$$O\left(\sqrt{\epsilon d} \log^{3/2}\left(\frac{1}{\epsilon d}\right) M \left\lceil \frac{T}{M} \right\rceil \right) = O\left(\sqrt{\epsilon d} \log^{3/2}\left(\frac{1}{\epsilon d}\right) T \right).$$

$\square$

Remark: The original work of [Koby Crammer and Vaughan, 2010] additionally allowed for some number K of "jumps": times t at which $\epsilon_t = 1$. Note that, in the above algorithm, since the influence of each sample is localized to the predictors trained within that "batch" of M instances, the effect of allowing such jumps would only change the bound on the number of mistakes to $\tilde{O}\left(\sqrt{d\epsilon}T + \sqrt{\frac{d}{\epsilon}}K\right)$. This compares favorably to the result of [Koby Crammer and Vaughan, 2010], which is roughly $O\left((d\epsilon)^{1/4}T + \frac{d^{1/4}}{\epsilon^{3/4}}K\right)$. However, the result of [Koby Crammer and Vaughan, 2010] was proven for a slightly more general setting, allowing distributions $\mathcal{P}$ that are not quite uniform (though they do require a relation between the angle between any two separators and the probability mass they disagree on, similar to that holding for the uniform distribution, which seems to require the distributions are not too far from uniform). It is not clear whether Theorem 11.8 can be generalized to this larger family of distributions.

11.6 General Analysis of Sublinear Mistake Bounds: Passive Learning

First, consider the following general lemma.

Lemma 11.16. *Suppose $\epsilon_t \to 0$. Then there exists an increasing sequence $\{T_i\}_{i=1}^\infty$ in $\mathbb{N}$ with $T_1 = 1$ such that $\lim_{i\to\infty} T_{i+1} - T_i = \infty$ while $\lim_{i\to\infty} \sum_{t=T_i}^{T_{i+1}-1} \epsilon_t = 0$.*

Proof. Let $T_1 = 1, T_2 = 2$, and $\gamma_2 = \epsilon_1$. Inductively, for each $i > 2$, if $\sum_{t=T_{i-1}}^{T_{i-1}+2(T_{i-1}-T_{i-2})-1} \epsilon_t \leq \gamma_{i-1}/2$, set $T_i = T_{i-1} + 2(T_{i-1} - T_{i-2})$ and $\gamma_i = \sum_{t=T_{i-1}}^{T_i-1} \epsilon_t$; otherwise, set $T_i = T_{i-1} + (T_{i-1} - T_{i-2})$ and $\gamma_i = \gamma_{i-1}$. Since any fixed value $k \in \mathbb{N}$ has $\lim_{T\to\infty} \sum_{t=T}^{T+k} \epsilon_t = 0$, we know there exist an infinite number of values $i \in \mathbb{N}$ with $\gamma_i \leq \gamma_{i-1}/2$, at which point we then also have $T_i - T_{i-1} = 2(T_{i-1} - T_{i-2}) > T_{i-1} - T_{i-2}$; together these facts imply the stated properties. $\square$

Suppose $\mathbb{C}$ is the concept space, and that $\mathbb{C}$ has finite VC dimension d. Consider the following passive learning algorithm, based on the sequnce T_i implied by Lemma 11.16.

> 0. Let $\hat{h}_1$ be any element of $\mathbb{C}$
>
> 1. For $i = 1, 2, \ldots$
>
> 2.   For $t = T_i, \ldots, T_{i+1} - 1$
>
> 3.   Predict $\hat{Y}_t = \hat{h}_i(X_t)$ as the prediction for the value of Y_t
>
> 4.   Let $\hat{h}_{i+1} = \mathrm{ERM}(\mathbb{C}, \{(X_{T_i}, Y_{T_i}), \ldots, (X_{T_{i+1}-1}, Y_{T_{i+1}-1})\})$

Theorem 11.17. *If $\epsilon_t \to 0$, and $\{T_i\}_{i=1}^{\infty}$ is the sequence guaranteed to exist by Lemma 11.16, then the above algorithm has an expected cumulative number of mistakes $o(T)$.*

Proof. Consider any value $i \in \mathbb{N}$, and let $h_{i+1} = h^*_{T_{i+1}}$. By a Chernoff bound, with probability at least $1 - 1/(T_{i+1} - T_i)$,

$$\sum_{t=T_i}^{T_{i+1}-1} \mathbb{I}[h_{i+1}(X_t) \neq h^*_t(X_t)] \leq \log_2(T_{i+1} - T_i) + 2e \sum_{t=T_i+1}^{T_{i+1}} \sum_{k=t}^{T_{i+1}} \epsilon_k.$$

Furthermore, standard VC analysis implies that, with probability at least $1 - 1/(T_{i+1} - T_i)$, $\forall h, g \in \mathbb{C}$,

$$\sum_{t=T_i}^{T_{i+1}-1} \mathbb{I}[h(X_t) \neq g(X_t)] \geq (T_{i+1} - T_i)\mathcal{P}(x : h(x) \neq g(x)) - c\sqrt{(d\log(T_{i+1} - T_i))(T_{i+1} - T_i)},$$

for some numerical constant $c > 0$. Thus, on these events, any $h \in \mathbb{C}$ with $\mathcal{P}(x : h(x) \neq h_{i+1}(x)) > 2\frac{\log_2(T_{i+1}-T_i)+2e\sum_{t=T_i+1}^{T_{i+1}}\sum_{k=t}^{T_{i+1}} \epsilon_k}{T_{i+1}-T_i} + c\sqrt{\frac{d\log(T_{i+1}-T_i)}{T_{i+1}-T_i}}$ must have

$$\sum_{t=T_i}^{T_{i+1}-1} \mathbb{I}[h(X_t) \neq h^*_t(X_t)]$$

$$\geq \sum_{t=T_i}^{T_{i+1}-1} \mathbb{I}[h(X_t) \neq h_{i+1}(X_t)] - \sum_{t=T_i}^{T_{i+1}-1} \mathbb{I}[h_{i+1}(X_t) \neq h^*_t(X_t)]$$

$$> \log_2(T_{i+1} - T_i) + 2e \sum_{t=T_i+1}^{T_{i+1}} \sum_{k=t}^{T_{i+1}} \epsilon_k$$

$$\geq \sum_{t=T_i}^{T_{i+1}-1} \mathbb{I}[h_{i+1}(X_t) \neq h^*_t(X_t)]$$

$$\geq \sum_{t=T_i}^{T_{i+1}-1} \mathbb{I}[\hat{h}_{i+1}(X_t) \neq h^*_t(X_t)].$$

Therefore, by a union bound, with probability at least $1 - 2/(T_{i+1} - T_i)$,

$$\mathcal{P}(x : \hat{h}_{i+1}(x) \neq h_{i+1}(x)) \leq 2\frac{\log_2(T_{i+1} - T_i) + 2e \sum_{t=T_i+1}^{T_{i+1}} \sum_{k=t}^{T_{i+1}} \epsilon_k}{T_{i+1} - T_i} + c\sqrt{\frac{d\log(T_{i+1} - T_i)}{T_{i+1} - T_i}},$$

so that

$$\mathbb{E}\left[\mathcal{P}(x : \hat{h}_{i+1}(x) \neq h_{i+1}(x))\right]$$
$$\leq 2\frac{\log_2(T_{i+1} - T_i) + 2e \sum_{t=T_i+1}^{T_{i+1}} \sum_{k=t}^{T_{i+1}} \epsilon_k}{T_{i+1} - T_i} + c\sqrt{\frac{d\log(T_{i+1} - T_i)}{T_{i+1} - T_i}} + \frac{2}{T_{i+1} - T_i}.$$

Denote by p_{i+1} the value on the right hand side of this inequality. Since $T_{i+1} - T_i \to \infty$ and $\frac{1}{T_{i+1} - T_i} \sum_{t=T_i+1}^{T_{i+1}} \sum_{k=t}^{T_{i+1}} \epsilon_k \leq \sum_{t=T_i+1}^{T_{i+1}} \epsilon_t \to 0$ (guaranteed by Lemma 11.16), we have $\lim_{i \to \infty} p_{i+1} = 0$. Since $\mathbb{E}[\sum_{t=T_{i+1}}^{T_{i+2}-1} \mathbb{I}[h_{i+1}(X_t) \neq h_t^*(X_t)]] \leq \sum_{t=T_{i+1}+1}^{T_{i+2}-1} \sum_{k=T_{i+1}+1}^{t} \epsilon_k$, we have

$$\mathbb{E}\left[\sum_{t=T_{i+1}}^{T_{i+2}-1} \mathbb{I}[\hat{h}_{i+1}(X_t) \neq h_t^*(X_t)]\right]$$
$$\leq \mathbb{E}\left[\sum_{t=T_{i+1}}^{T_{i+2}-1} \mathbb{I}[\hat{h}_{i+1}(X_t) \neq h_{i+1}(X_t)]\right] + \mathbb{E}\left[\sum_{t=T_{i+1}}^{T_{i+2}-1} \mathbb{I}[h_{i+1}(X_t) \neq h_t^*(X_t)]\right]$$
$$\leq (T_{i+2} - T_{i+1})\mathbb{E}[\mathcal{P}(x : \hat{h}_{i+1}(x) \neq h_{i+1}(x))] + \sum_{t=T_{i+1}+1}^{T_{i+2}-1} \sum_{k=T_{i+1}+1}^{t} \epsilon_k$$
$$\leq (T_{i+2} - T_{i+1})p_{i+1} + \sum_{t=T_{i+1}+1}^{T_{i+2}-1} \sum_{k=T_{i+1}+1}^{t} \epsilon_k.$$

Since $p_{i+1} \to 0$, we have $(T_{i+2} - T_{i+1})p_{i+1} = o(T_{i+2} - T_{i+1})$, and since $T_{i+2} - T_{i+1} \to \infty$, we have $\sum_{i=1}^{j}(T_{i+2} - T_{i+1})p_{i+1} = o(T_j)$. Furthermore, since $\sum_{t=T_{i+1}+1}^{T_{i+2}-1} \sum_{k=T_{i+1}+1}^{t} \epsilon_k \leq (T_{i+2} - T_{i+1}) \sum_{t=T_{i+1}+1}^{T_{i+2}-1} \epsilon_t = o(T_{i+2} - T_{i+1})$, and $T_{i+2} - T_{i+1} \to \infty$, we have $\sum_{i=1}^{j} \sum_{t=T_{i+1}+1}^{T_{i+2}-1} \sum_{k=T_{i+1}+1}^{t} \epsilon_k = o(T_j)$. Altogether, we have that the expected sum of mistakes up to time T (which is the sum of the expected numbers of mistakes within the component segments $T_{i+1}, \ldots, T_{i+2} - 1$) grows sublinearly in T. $\qquad \square$

11.7 General Analysis under Varying Drift Rate: Inefficient Passive Learning

Consider the following algorithm.

---

0. For $T = 1, 2, \ldots$

1. Let $m_T = \mathrm{argmin}_{m \in \{1,\ldots,T-1\}} \sum_{t=T-m+1}^{T} \epsilon_t + \frac{d \log(m/d)}{m}$

2. Let $\hat{h}_T = \mathrm{ERM}(\mathbb{C}, \{(X_{T-m_T}, Y_{T-m_T}), \ldots, (X_{T-1}, Y_{T-1})\})$

3. Predict $\hat{Y}_T = \hat{h}_T(X_T)$ as the prediction for the value of Y_T

Theorem 11.18. *The above algorithm makes an expected number of mistakes among the first T instances that is*

$$O \left(\sum_{t=1}^{T} \min_{m \in \{1,\ldots,t-1\}} \sum_{s=t-m+1}^{t} \epsilon_s + \frac{d \log(m/d)}{m} \right).$$

Proof. It suffices to show that, for any $T \in \mathbb{N}$, and any $m \in \{1, \ldots, T-1\}$, the classifier $\hat{h} = \mathrm{ERM}(\mathbb{C}, \{(X_{T-m}, Y_{T-m}), \ldots, (X_{T-1}, Y_{T-1})\})$ has

$$\mathbb{E}[\mathcal{P}(x : \hat{h}(x) \neq h_T^*(x))] \leq c' \left(\sum_{t=T-m+1}^{T} \epsilon_t + \frac{d \log(m/d)}{m} \right),$$

for some universal constant $c' \in (0, \infty)$. Minimization over m in the theorem statement then follows from the fact that m_T minimizes this expression over m by definition. The result will then follow by linearity of expectations.

Let $\mathcal{E} = \sum_{t=T-m+1}^{T} \epsilon_t$. By a Chernoff bound, with probability at least $1 - \delta$,

$$\frac{1}{m} \sum_{i=T-m}^{T-1} \mathbb{I}[h_{T-m}^*(X_i) \neq h_i^*(X_i)] \leq \frac{\log_2(1/\delta) + 2em\mathcal{E}}{m} = \frac{\log_2(1/\delta)}{m} + 2e\mathcal{E}.$$

In particular, this means

$$\frac{1}{m} \sum_{i=T-m}^{T-1} \mathbb{I}[\hat{h}(X_i) \neq h_{T-m}^*(X_i)] \leq \frac{2 \log_2(1/\delta)}{m} + 4e\mathcal{E}.$$

214

By Lemma 11.1, on an additional event of probability at least $1 - \delta$,

$$
\mathcal{P}(x : \hat{h}(x) \neq h^*_{T-m}(x))
$$

$$
\leq \frac{2 \log_2(1/\delta)}{m} + 4e\mathcal{E} + c\sqrt{\left(\frac{2 \log_2(1/\delta)}{m} + 4e\mathcal{E}\right) \frac{d \log(m/d) + \log(1/\delta)}{m}} + c\frac{d \log(m/d) + \log(1/\delta)}{m}
$$

$$
\leq c'' \left(\mathcal{E} + \sqrt{\mathcal{E} \frac{d \log(m/d) + \log(1/\delta)}{m}} + \frac{d \log(m/d) + \log(1/\delta)}{m} \right),
$$

for an appropriate numerical constant $c'' \in [1, \infty)$. Taking $\delta = d/m$, this is at most

$$
2c'' \left(\mathcal{E} + \sqrt{\mathcal{E} \frac{d \log(m/d)}{m}} + \frac{d \log(m/d)}{m} \right).
$$

Since this holds with probability $1 - 2\delta = 1 - 2d/m$, and $\mathcal{P}(x : \hat{h}(x) \neq h^*_{T-m}(x)) \leq 1$ always,

we have

$$
\mathbb{E}\left[\mathcal{P}(x : \hat{h}(x) \neq h^*_T(x))\right] \leq \mathbb{E}\left[\mathcal{P}(x : \hat{h}(x) \neq h^*_{T-m}(x))\right] + \mathcal{P}(x : h^*_{T-m}(x) \neq h^*_T(x))
$$

$$
\leq 2c'' \left(\mathcal{E} + \sqrt{\mathcal{E} \frac{d \log(m/d)}{m}} + \frac{d \log(m/d)}{m} \right) + 2\frac{d}{m} + \mathcal{E}
$$

$$
\leq 4c'' \left(\mathcal{E} + \sqrt{\mathcal{E} \frac{d \log(m/d)}{m}} + \frac{d \log(m/d)}{m} \right)
$$

$$
\leq 4c'' \left(\sqrt{\mathcal{E}} + \sqrt{\frac{d \log(m/d)}{m}} \right)^2
$$

$$
\leq 16c'' \max\left\{ \mathcal{E}, \frac{d \log(m/d)}{m} \right\}
$$

$$
\leq 16c'' \left(\mathcal{E} + \frac{d \log(m/d)}{m} \right).
$$

$\square$

In particular, we have the following corollary.

Corollary 11.19. *If $\sum_{t=1}^T \epsilon_t = o(T)$, then the expected number of mistakes made by the above algorithm is also $o(T)$.*

Proof. Let $\beta_t(m) = \max\left\{\sum_{s=t-m+1}^{t} \epsilon_s, \frac{d\log(m/d)}{m}\right\}$, and note that

$$\sum_{s=t-m+1}^{t} \epsilon_s \quad + \quad \frac{d\log(m/d)}{m} \quad \leq \quad 2\beta_t(m),$$

so that Theorem 11.18 (combined with the fact that the probability of a mistake on a given round is at most 1) implies the expected number of mistakes is $O(\sum_{t=1}^{T} \min_{m\in\{1,\ldots,t-1\}} \beta_t(m) \wedge 1)$. Let $m_t' = \operatorname{argmin}_{m\in\{1,\ldots,t-1\}} \beta_t(m)$.

Fix any $M \in \mathbb{N}$. For a given t, if $m_t' < M$, then it must be that $\sum_{s=t-M+1}^{t} \epsilon_s > \frac{d\log(M/d)}{M}$. Also, since

$$\sum_{t=M}^{T} \sum_{s=t-M+1}^{t} \epsilon_s = \sum_{t=1}^{M-1} t\epsilon_t + M\sum_{t=M}^{T} \epsilon_t = o(T),$$

and

$$\sum_{t=M}^{T} \mathbb{I}\left[\sum_{s=t-M+1}^{t} \epsilon_s > \frac{d\log(M/d)}{M}\right] \leq \frac{M}{d\log(M/d)} \sum_{t=M}^{T} \sum_{s=t-M+1}^{t} \epsilon_s,$$

we have that

$$\sum_{t=M}^{T} \mathbb{I}\left[m_t' < M\right] = o(T).$$

Furthermore, consider any t for which $m_t' \geq M$. Then

$$\min_{m\in\{1,\ldots,t-1\}} \beta_t(m) \leq \max\left\{\sum_{s=t-M+1}^{t} \epsilon_s, \frac{d\log(M/d)}{M}\right\}.$$

As established above,

$$\sum_{t=M}^{T} \mathbb{I}\left[\sum_{s=t-M+1}^{t} \epsilon_s > \frac{d\log(M/d)}{M}\right] = o(T),$$

so that

$$\sum_{t=1}^{T} \min_{m\in\{1,\ldots,t-1\}} \beta_t(m) \wedge 1$$

$$\leq \frac{d\log(M/d)}{M}T + \sum_{t=1}^{T} \mathbb{I}\left[m_t' < M \text{ or } m_t' \geq M \text{ and } \sum_{s=t-M+1}^{t} \epsilon_s > \frac{d\log(M/d)}{M}\right]$$

$$\leq \frac{d\log(M/d)}{M}T + \sum_{t=1}^{T} \mathbb{I}\left[\sum_{s=t-M+1}^{t} \epsilon_s > \frac{d\log(M/d)}{M}\right]$$

$$= \frac{d\log(M/d)}{M}T + o(T).$$

216

Since this is true of any $M \in \mathbb{N}$, we have that

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \min_{m \in \{1,\ldots,t-1\}} \beta_t(m) \wedge 1 \leq \lim_{M \to \infty} \frac{d \log(M/d)}{M} = 0,$$

so that the expected number of mistakes is $o(T)$, as claimed. $\qquad \square$

Chapter 12

Surrogate Losses in Passive and Active Learning

Abstract

[1] Active learning is a type of sequential design for supervised machine learning, in which the learning algorithm sequentially requests the labels of selected instances from a large pool of unlabeled data points. The objective is to produce a classifier of relatively low risk, as measured under the 0-1 loss, ideally using fewer label requests than the number of random labeled data points sufficient to achieve the same. This work investigates the potential uses of surrogate loss functions in the context of active learning. Specifically, it presents an active learning algorithm based on an arbitrary classification-calibrated surrogate loss function, along with an analysis of the number of label requests sufficient for the classifier returned by the algorithm to achieve a given risk under the 0-1 loss. Interestingly, these results cannot be obtained by simply optimizing the surrogate risk via active learning to an extent sufficient to provide a guarantee on the 0-1 loss, as is common practice in the analysis of surrogate losses for passive learning. Some of the results have additional implications for the use of surrogate losses in passive learning.

---

[1]The chapter is based on joint work with Steve Hanneke.

12.1 Introduction

In supervised machine learning, we are tasked with learning a classifier whose probability of making a mistake (i.e., error rate) is small. The study of when it is possible to learn an accurate classifier via a computationally efficient algorithm, and how to go about doing so, is a subtle and difficult topic, owing largely to nonconvexity of the loss function: namely, the 0-1 loss. While there is certainly an active literature on developing computationally efficient methods that succeed at this task, even under various noise conditions, it seems fair to say that at present, many of these advances have not yet reached the level of robustness, efficiency, and simplicity required for most applications. In the mean time, practitioners have turned to various heuristics in the design of practical learning methods, in attempts to circumvent these tough computational problems. One of the most common such heuristics is the use of a convex *surrogate* loss function in place of the 0-1 loss in various optimizations performed by the learning method. The convexity of the surrogate loss allows these optimizations to be performed efficiently, so that the methods can be applied within a reasonable execution time, even with only modest computational resources. Although classifiers arrived at in this way are not always guaranteed to be good classifiers when performance is measured under the 0-1 loss, in practice this heuristic has often proven quite effective. In light of this fact, most modern learning methods either explicitly make use of a surrogate loss in the formulation of optimization problems (e.g., SVM), or implicitly optimize a surrogate loss via iterative descent (e.g., AdaBoost). Indeed, the choice of a surrogate loss is often as fundamental a part of the process of approaching a learning problem as the choice of hypothesis class or learning bias. Thus it seems essential that we come to some understanding of how best to make use of surrogate losses in the design of learning methods, so that in the favorable scenario that this heuristic actually does work, we have methods taking full advantage of it.

In this work, we are primarily interested in how best to use surrogate losses in the context of *active learning*, which is a type of sequential design in which the learning algorithm is pre-

sented with a large pool of unlabeled data points (i.e., only the covariates are observable), and can sequentially request to observe the labels (response variables) of individual instances from the pool. The objective in active learning is to produce a classifier of low error rate while accessing a smaller number of labels than would be required for a method based on random labeled data points (i.e., *passive learning*) to achieve the same. We take as our starting point that we have already committed to use a given surrogate loss, and we restrict our attention to just those scenarios in which this heuristic actually *does* work. We are then interested in how best to make use of the surrogate loss toward the goal of producing a classifier with relatively small error rate. To be clear, we focus on the case where the minimizer of the surrogate risk also minimizes the error rate, and is contained in our function class.

We construct an active learning strategy based on optimizing the empirical surrogate risk over increasingly focused subsets of the instance space, and derive bounds on the number of label requests the method requires to achieve a given error rate. Interestingly, we find that the basic approach of optimizing the surrogate risk via active learning to a sufficient extent to guarantee small error rate generally does not lead to as strong of results. In fact, the method our results apply to typically *does not* optimize the surrogate risk (even in the limit). The insight leading to this algorithm is that, if we are truly only interested in achieving low 0-1 loss, then once we have identified the *sign* of the optimal function at a given point, we need not optimize the value of the function at that point any further, and can therefore focus the label requests elsewhere. As a byproduct of this analysis, we find this insight has implications for the use of certain surrogate losses in passive learning as well, though to a lesser extent.

Most of the mathematical tools used in this analysis are inspired by recently-developed techniques for the study of active learning [Hanneke, 2009, 2011, Koltchinskii, 2010], in conjunction with the results of Bartlett, Jordan, and McAuliffe [2006] bounding the excess error rate in terms of the excess surrogate risk, and the works of Koltchinskii [2006] and Bartlett, Bousquet, and Mendelson [2005] on localized Rademacher complexity bounds.

12.1.1 Related Work

There are many previous works on the topic of surrogate losses in the context of passive learning. Perhaps the most relevant to our results below are the work of Bartlett, Jordan, and McAuliffe [2006] and the related work of Zhang [2004]. These develop a general theory for converting results on excess risk under the surrogate loss into results on excess risk under the 0-1 loss. Below, we describe the conclusions of that work in detail, and we build on many of the basic definitions and insights pioneered in these works.

Another related line of research, initiated by Audibert and Tsybakov [2007], studies "plug-in rules," which make use of regression estimates obtained by optimizing a surrogate loss, and are then rounded to $\{-1, +1\}$ values to obtain classifiers. They prove results under smoothness assumptions on the actual regression function, which (remarkably) are often *better* than the known results for methods that directly optimize the 0-1 loss. Under similar conditions, Minsker [2012] studies an analogous active learning method, which again makes use of a surrogate loss, and obtains improvements in label complexity compared to the passive learning method of Audibert and Tsybakov [2007]; again, the results for this method based on a surrogate loss are actually better than those derived from existing active learning methods designed to directly optimize the 0-1 loss. The works of Audibert and Tsybakov [2007] and Minsker [2012] raise interesting questions about whether the general analyses of methods that optimize the 0-1 loss remain tight under complexity assumptions on the regression function, and potentially also about the design of optimal methods for classification when assumptions are phrased in terms of the regression function.

In the present work, we focus our attention on scenarios where the main purpose of using the surrogate loss is to ease the computational problems associated with minimizing an empirical risk, so that our statistical results are typically strongest when the surrogate loss is the 0-1 loss itself. Thus, in the specific scenarios studied by Minsker [2012], our results are generally not optimal; rather, the main strength of our analysis lies in its generality. In this sense, our results

are more closely related to those of Bartlett, Jordan, and McAuliffe [2006] and Zhang [2004] than to those of Audibert and Tsybakov [2007] and Minsker [2012]. That said, we note that several important elements of the design and analysis of the active learning method below are already present to some extent in the work of Minsker [2012].

There are several interesting works on active learning methods that optimize a general loss function. Beygelzimer, Dasgupta, and Langford [2009] and Koltchinskii [2010] have both proposed active learning methods, and analyzed the number of label requests the methods make before achieving a given excess risk for that loss function. The former method is based on importance weighted sampling, while the latter makes clear an interesting connection to local Rademacher complexities. One natural idea for approaching the problem of active learning with a surrogate loss is to run one of these methods with the surrogate loss. The results of Bartlett, Jordan, and McAuliffe [2006] allow us to determine a sufficiently small value γ such that any function with excess surrogate risk at most γ has excess error rate at most ε. Thus, by evaluating the established bounds on the number of label requests sufficient for these active learning methods to achieve excess surrogate risk γ, we immediately have a result on the number of label requests sufficient for them to achieve excess error rate ε. This is a common strategy for constructing and analyzing passive learning algorithms that make use of a surrogate loss. However, as we discuss below, this strategy does not generally lead to the best behavior in active learning, and often will not be much better than simply using a related passive learning method. Instead, we propose a new method that typically does not optimize the surrogate risk, but makes use of it in a different way so as to achieve stronger results when performance is measured under the 0-1 loss.

12.2 Definitions

Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be a measurable space, where $\mathcal{X}$ is called the *instance space*; for convenience, we suppose this is a standard Borel space. Let $\mathcal{Y} = \{-1, +1\}$, and equip the space $\mathcal{X} \times \mathcal{Y}$ with its

product σ-algebra: $\mathcal{B} = \mathcal{B}_{\mathcal{X}} \otimes 2^{\mathcal{Y}}$. Let $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, let $\mathcal{F}^*$ denote the set of all measurable functions $g : \mathcal{X} \to \bar{\mathbb{R}}$, and let $\mathcal{F} \subseteq \mathcal{F}^*$, where $\mathcal{F}$ is called the *function class*. Throughout, we fix a distribution $\mathcal{P}_{XY}$ over $\mathcal{X} \times \mathcal{Y}$, and we denote by $\mathcal{P}$ the marginal distribution of $\mathcal{P}_{XY}$ over $\mathcal{X}$. In the analysis below, we make the usual simplifying assumption that the events and functions in the definitions and proofs are indeed measurable. In most cases, this holds under simple conditions on $\mathcal{F}$ and $\mathcal{P}_{XY}$ [see e.g., van der Vaart and Wellner, 2011]; when this is not the case, we may turn to outer probabilities. However, we will not discuss these technical issues further.

For any $h \in \mathcal{F}^*$, and any distribution P over $\mathcal{X} \times \mathcal{Y}$, denote the *error rate* by $\mathrm{er}(h; P) = P((x, y) : \mathrm{sign}(h(x)) \neq y)$; when $P = \mathcal{P}_{XY}$, we abbreviate this as $\mathrm{er}(h) = \mathrm{er}(h; \mathcal{P}_{XY})$. Also, let $\eta(X; P)$ be a version of $\mathbb{P}(Y = 1|X)$, for $(X, Y) \sim P$; when $P = \mathcal{P}_{XY}$, abbreviate this as $\eta(X) = \eta(X; \mathcal{P}_{XY})$. In particular, note that $\mathrm{er}(h; P)$ is minimized at any h with $\mathrm{sign}(h(x)) = \mathrm{sign}(\eta(x; P) - 1/2)$ for all $x \in \mathcal{X}$. In this work, we will also be interested in certain conditional distributions and modifications of functions, specified as follows. For any measurable $\mathcal{U} \subseteq \mathcal{X}$ with $\mathcal{P}(\mathcal{U}) > 0$, define the probability measure $\mathcal{P}_{\mathcal{U}}(\cdot) = \mathcal{P}_{XY}(\cdot|\mathcal{U} \times \mathcal{Y}) = \mathcal{P}_{XY}(\cdot \cap \mathcal{U} \times \mathcal{Y})/\mathcal{P}(\mathcal{U})$: that is, $\mathcal{P}_{\mathcal{U}}$ is the conditional distribution of $(X, Y) \sim \mathcal{P}_{XY}$ given that $X \in \mathcal{U}$. Also, for any $h, g \in \mathcal{F}^*$, define the spliced function $h_{\mathcal{U},g}(x) = h(x)\mathbb{I}_{\mathcal{U}}(x) + g(x)\mathbb{I}_{\mathcal{X} \setminus \mathcal{U}}(x)$. For a set $\mathcal{H} \subseteq \mathcal{F}^*$, denote $\mathcal{H}_{\mathcal{U},g} = \{h_{\mathcal{U},g} : h \in \mathcal{H}\}$.

For any $\mathcal{H} \subseteq \mathcal{F}^*$, define the *region of sign-disagreement* $\mathrm{DIS}(\mathcal{H}) = \{x \in \mathcal{X} : \exists h, g \in \mathcal{H} \text{ s.t. } \mathrm{sign}(h(x)) \neq \mathrm{sign}(g(x))\}$, and the *region of value-disagreement* $\mathrm{DISF}(\mathcal{H}) = \{x \in \mathcal{X} : \exists h, g \in \mathcal{H} \text{ s.t. } h(x) \neq g(x)\}$, and denote by $\overline{\mathrm{DIS}}(\mathcal{H}) = \mathrm{DIS}(\mathcal{H}) \times \mathcal{Y}$ and $\overline{\mathrm{DISF}}(\mathcal{H}) = \mathrm{DISF}(\mathcal{H}) \times \mathcal{Y}$. Additionally, we denote by $[\mathcal{H}] = \{f \in \mathcal{F}^* : \forall x \in \mathcal{X}, \inf_{h \in \mathcal{H}} h(x) \leq f(x) \leq \sup_{h \in \mathcal{H}} h(x)\}$ the minimal bracket set containing $\mathcal{H}$.

Our interest here is learning from data, so let $\mathcal{Z} = \{(X_1, Y_1), (X_2, Y_2), \ldots\}$ denote a sequence of independent $\mathcal{P}_{XY}$-distributed random variables, referred to as the *labeled data* sequence, while $\{X_1, X_2, \ldots\}$ is referred to as the *unlabeled data* sequence. For $m \in \mathbb{N}$, we also denote $\mathcal{Z}_m = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$. Throughout, we will let $\delta \in (0, 1/4)$ denote an arbitrary confidence

parameter, which will be referenced in the methods and theorem statements.

The *active learning* protocol is defined as follows. An active learning algorithm is initially permitted access to the sequence $X_1, X_2, \ldots$ of unlabeled data. It may then select an index $i_1 \in \mathbb{N}$ and *request* to observe Y_{i_1}; after observing Y_{i_1}, it may select another index $i_2 \in \mathbb{N}$, request to observe Y_{i_2}, and so on. After a number of such label requests not exceeding some specified budget n, the algorithm halts and returns a function $\hat{h} \in \mathcal{F}^*$. Formally, this protocol specifies a type of mapping that maps the random variable $\mathcal{Z}$ to a function $\hat{h}$, where $\hat{h}$ is conditionally independent of $\mathcal{Z}$ given $X_1, X_2, \ldots$ and $(i_1, Y_{i_1}), (i_2, Y_{i_2}), \ldots, (i_n, Y_{i_n})$, where each i_k is conditionally independent of $\mathcal{Z}$ and $i_{k+1}, \ldots, i_n$ given $X_1, X_2, \ldots$ and $(i_1, Y_{i_1}), \ldots, (i_{k-1}, Y_{i_{k-1}})$.

12.2.1 Surrogate Loss Functions for Classification

Throughout, we let $\ell : \bar{\mathbb{R}} \to [0, \infty]$ denote an arbitrary *surrogate loss function*; we will primarily be interested in functions ℓ that satisfy certain conditions discussed below. To simplify some statements below, it will be convenient to suppose $z \in \mathbb{R} \Rightarrow \ell(z) < \infty$. For any $g \in \mathcal{F}^*$ and distribution P over $\mathcal{X} \times \mathcal{Y}$, let $\mathrm{R}_\ell(g; P) = \mathbb{E}[\ell(g(X)Y)]$, where $(X, Y) \sim P$; in the case $P = \mathcal{P}_{XY}$, abbreviate $\mathrm{R}_\ell(g) = \mathrm{R}_\ell(g; \mathcal{P}_{XY})$. Also define $\bar{\ell} = 1 \vee \sup_{x \in \mathcal{X}} \sup_{h \in \mathcal{F}} \max_{y \in \{-1, +1\}} \ell(yh(x))$; we will generally suppose $\bar{\ell} < \infty$. In practice, this is more often a constraint on $\mathcal{F}$ than on ℓ; that is, we could have ℓ unbounded, but due to some normalization of the functions $h \in \mathcal{F}$, ℓ is bounded on the corresponding set of values.

Throughout this work, we will be interested in loss functions ℓ whose point-wise minimizer necessarily also optimizes the 0-1 loss. This property was nicely characterized by Bartlett, Jordan, and McAuliffe [2006] as follows. For $\eta_0 \in [0, 1]$, define $\ell^\star(\eta_0) = \inf_{z \in \mathbb{R}}(\eta_0 \ell(z) + (1 - \eta_0)\ell(-z))$, and $\ell_-^\star(\eta_0) = \inf_{z \in \mathbb{R} : z(2\eta_0 - 1) \leq 0}(\eta_0 \ell(z) + (1 - \eta_0)\ell(-z))$.

Definition 12.1. *The loss ℓ is* classification-calibrated *if,* $\forall \eta_0 \in [0, 1] \setminus \{1/2\}$, $\ell_-^\star(\eta_0) > \ell^\star(\eta_0)$.

In our context, for $X \sim \mathcal{P}$, $\ell^\star(\eta(X))$ represents the minimum value of the conditional ℓ-risk at X, so that $\mathbb{E}[\ell^\star(\eta(X))] = \inf_{h \in \mathcal{F}^*} \mathrm{R}_\ell(h)$, while $\ell_-^\star(\eta(X))$ represents the minimum conditional

224

ℓ-risk at X, subject to having a sub-optimal conditional error rate at X: i.e., $\mathrm{sign}(h(X)) \neq \mathrm{sign}(\eta(X) - 1/2)$. Thus, being classification-calibrated implies the minimizer of the conditional ℓ-risk at X necessarily has the same sign as the minimizer of the conditional error rate at X. Since we are only interested here in using ℓ as a reasonable surrogate for the 0-1 loss, throughout the work below we suppose ℓ is classification-calibrated.

Though not strictly necessary for our results below, it will be convenient for us to suppose that, for all $\eta_0 \in [0, 1]$, this infimum value $\ell^\star(\eta_0)$ is actually *obtained* as $\eta_0 \ell(z^\star(\eta_0)) + (1 - \eta_0)\ell(-z^\star(\eta_0))$ for some $z^\star(\eta_0) \in \bar{\mathbb{R}}$ (not necessarily unique). For instance, this is the case for any nonincreasing right-continuous ℓ, or continuous and convex ℓ, which include most of the cases we are interested in using as surrogate losses anyway. The proofs can be modified in a natural way to handle the general case, simply substituting any z with conditional risk sufficiently close to the minimum value. For any distribution P, denote $h^*_P(x) = z^\star(\eta(x; P))$ for all $x \in \mathcal{X}$. In particular, note that h^*_P obtains $\mathrm{R}_\ell(h^*_P; P) = \inf_{g \in \mathcal{F}^*} \mathrm{R}_\ell(g; P)$. When $P = \mathcal{P}_{XY}$, we abbreviate this as $h^* = h^*_{\mathcal{P}_{XY}}$. Furthermore, if ℓ is classification-calibrated, then $\mathrm{sign}(h^*_P(x)) = \mathrm{sign}(\eta(x; P) - 1/2)$ for all $x \in \mathcal{X}$ with $\eta(x; P) \neq 1/2$, and hence $\mathrm{er}(h^*_P; P) = \inf_{h \in \mathcal{F}^*} \mathrm{er}(h; P)$ as well.

For any distribution P over $\mathcal{X} \times \mathcal{Y}$, and any $h, g \in \mathcal{F}^*$, define the *loss distance* $\mathrm{D}_\ell(h, g; P) = \sqrt{\mathbb{E}\left[(\ell(h(X)Y) - \ell(g(X)Y))^2\right]}$, where $(X, Y) \sim P$. Also define the *loss diameter* of a class $\mathcal{H} \subseteq \mathcal{F}^*$ as $\mathrm{D}_\ell(\mathcal{H}; P) = \sup_{h,g \in \mathcal{H}} \mathrm{D}_\ell(h, g; P)$, and the ℓ-risk ε-minimal set of $\mathcal{H}$ as $\mathcal{H}(\varepsilon; \ell, P) = \{h \in \mathcal{H} : \mathrm{R}_\ell(h; P) - \inf_{g \in \mathcal{H}} \mathrm{R}_\ell(g; P) \leq \varepsilon\}$. When $P = \mathcal{P}_{XY}$, we abbreviate these as $\mathrm{D}_\ell(h, g) = \mathrm{D}_\ell(h, g; \mathcal{P}_{XY})$, $\mathrm{D}_\ell(\mathcal{H}) = \mathrm{D}_\ell(\mathcal{H}; \mathcal{P}_{XY})$, and $\mathcal{H}(\varepsilon; \ell) = \mathcal{H}(\varepsilon; \ell, \mathcal{P}_{XY})$. Also, for any $h \in \mathcal{F}^*$, abbreviate $h_\mathcal{U} = h_{\mathcal{U}, h^*}$, and for any $\mathcal{H} \subseteq \mathcal{F}^*$, define $\mathcal{H}_\mathcal{U} = \{h_\mathcal{U} : h \in \mathcal{H}\}$.

We additionally define related quantities for the 0-1 loss, as follows. Define the *distance* $\Delta_P(h, g) = \mathcal{P}(x : \mathrm{sign}(h(x)) \neq \mathrm{sign}(g(x)))$ and *radius* $\mathrm{radius}(\mathcal{H}; P) = \sup_{h \in \mathcal{H}} \Delta_P(h, h^*_P)$. Also define the ε-minimal set of $\mathcal{H}$ as $\mathcal{H}(\varepsilon; 01, P) = \{h \in \mathcal{H} : \mathrm{er}(h; P) - \inf_{g \in \mathcal{H}} \mathrm{er}(g; P) \leq \varepsilon\}$, and for $r > 0$, define the r-ball centered at h in $\mathcal{H}$ by $\mathrm{B}_{\mathcal{H}, P}(h, r) = \{g \in \mathcal{H} : \Delta_P(h, g) \leq r\}$.

When $P = \mathcal{P}_{XY}$, we abbreviate these as $\Delta(h, g) = \Delta_{\mathcal{P}_{XY}}(h, g)$, $\mathrm{radius}(\mathcal{H}) = \mathrm{radius}(\mathcal{H}; \mathcal{P}_{XY})$, $\mathcal{H}(\varepsilon; 01) = \mathcal{H}(\varepsilon; 01, \mathcal{P}_{XY})$, and $\mathrm{B}_{\mathcal{H}}(h, r) = \mathrm{B}_{\mathcal{H}, \mathcal{P}_{XY}}(h, r)$; when $\mathcal{H} = \mathcal{F}$, further abbreviate $\mathrm{B}(h, r) = \mathrm{B}_{\mathcal{F}}(h, r)$.

We will be interested in transforming results concerning the excess surrogate risk into results on the excess error rate. As such, we will make use of the following abstract transformation.

Definition 12.2. *For any distribution P over $\mathcal{X} \times \mathcal{Y}$, and any $\varepsilon \in [0, 1]$, define*

$$\Gamma_\ell(\varepsilon; P) = \sup\{\gamma > 0 : \mathcal{F}^*(\gamma; \ell, P) \subseteq \mathcal{F}^*(\varepsilon; 01, P)\} \cup \{0\}.$$

Also, for any $\gamma \in [0, \infty)$, define the inverse

$$\mathcal{E}_\ell(\gamma; P) = \inf \{\varepsilon > 0 : \gamma \leq \Gamma_\ell(\varepsilon; P)\}.$$

When $P = \mathcal{P}_{XY}$, abbreviate $\Gamma_\ell(\varepsilon) = \Gamma_\ell(\varepsilon; \mathcal{P}_{XY})$ and $\mathcal{E}_\ell(\gamma) = \mathcal{E}_\ell(\gamma; \mathcal{P}_{XY})$.

By definition, Γ_ℓ has the property that

$$\forall h \in \mathcal{F}^*, \forall \varepsilon \in [0, 1], \ \ \mathrm{R}_\ell(h) - \mathrm{R}_\ell(h^*) < \Gamma_\ell(\varepsilon) \implies \mathrm{er}(h) - \mathrm{er}(h^*) \leq \varepsilon. \tag{12.1}$$

In fact, Γ_ℓ is defined to be maximal with this property, in that *any* Γ'_ℓ for which (12.1) is satisfied must have $\Gamma'_\ell(\varepsilon) \leq \Gamma_\ell(\varepsilon)$ for all $\varepsilon \in [0, 1]$.

In our context, we will typically be interested in calculating lower bounds on Γ_ℓ for any particular scenario of interest. Bartlett, Jordan, and McAuliffe [2006] studied various lower bounds of this type. Specifically, for $\zeta \in [-1, 1]$, define $\tilde{\psi}_\ell(\zeta) = \ell_-^\star\left(\frac{1+\zeta}{2}\right) - \ell^\star\left(\frac{1+\zeta}{2}\right)$, and let ψ_ℓ be the largest convex lower bound of $\tilde{\psi}_\ell$ on $[0, 1]$, which is well-defined in this context [Bartlett, Jordan, and McAuliffe, 2006]; for convenience, also define $\psi_\ell(x)$ for $x \in (1, \infty)$ arbitrarily subject to maintaining convexity of ψ_ℓ. Bartlett, Jordan, and McAuliffe [2006] show ψ_ℓ is continuous and nondecreasing on $(0, 1)$, and in fact that $x \mapsto \psi_\ell(x) / x$ is nondecreasing on $(0, \infty)$. They also show every $h \in \mathcal{F}^*$ has $\psi_\ell(\mathrm{er}(h) - \mathrm{er}(h^*)) \leq \mathrm{R}_\ell(h) - \mathrm{R}_\ell(h^*)$, so that $\psi_\ell \leq \Gamma_\ell$, and they find this inequality can be tight for a particular choice of $\mathcal{P}_{XY}$. They further study more subtle relationships between excess ℓ-risk and excess error rate holding for any classification-calibrated ℓ. In particular, following the same argument as in the proof of their Theorem 3, one

can show that if ℓ is classification-calibrated, every $h \in \mathcal{F}^*$ satisfies

$$\Delta(h, h^*) \cdot \psi_\ell \left(\frac{\mathrm{er}(h) - \mathrm{er}(h^*)}{2\Delta(h, h^*)} \right) \leq \mathrm{R}_\ell(h) - \mathrm{R}_\ell(h^*).$$

The implication of this in our context is the following. Fix any nondecreasing function $\Psi_\ell :$ $[0,1] \to [0,\infty)$ such that $\forall \varepsilon \geq 0$,

$$\Psi_\ell(\varepsilon) \leq \mathrm{radius}(\mathcal{F}^*(\varepsilon; \text{01}))\psi_\ell \left(\frac{\varepsilon}{2\mathrm{radius}(\mathcal{F}^*(\varepsilon; \text{01}))} \right). \tag{12.2}$$

Any $h \in \mathcal{F}^*$ with $\mathrm{R}_\ell(h) - \mathrm{R}_\ell(h^*) < \Psi_\ell(\varepsilon)$ also has $\Delta(h, h^*)\psi_\ell \left(\frac{\mathrm{er}(h) - \mathrm{er}(h^*)}{2\Delta(h,h^*)} \right) < \Psi_\ell(\varepsilon)$; combined with the fact that $x \mapsto \psi_\ell(x)/x$ is nondecreasing on $(0,1)$, this implies $\mathrm{radius}(\mathcal{F}^*(\mathrm{er}(h) - \mathrm{er}(h^*); \text{01}))\psi_\ell \left(\frac{\mathrm{er}(h) - \mathrm{er}(h^*)}{2\mathrm{radius}(\mathcal{F}^*(\mathrm{er}(h) - \mathrm{er}(h^*);\text{01}))} \right) < \Psi_\ell(\varepsilon)$; this means $\Psi_\ell(\mathrm{er}(h) - \mathrm{er}(h^*)) < \Psi_\ell(\varepsilon)$, and monotonicity of Ψ_ℓ implies $\mathrm{er}(h) - \mathrm{er}(h^*) < \varepsilon$. Altogether, this implies $\Psi_\ell(\varepsilon) \leq \Gamma_\ell(\varepsilon)$. In fact, though we do not present the details here, with only minor modifications to the proofs below, when $h^* \in \mathcal{F}$, all of our results involving $\Gamma_\ell(\varepsilon)$ will also hold while replacing $\Gamma_\ell(\varepsilon)$ with any nondecreasing Ψ_ℓ' such that $\forall \varepsilon \geq 0$,

$$\Psi_\ell'(\varepsilon) \leq \mathrm{radius}(\mathcal{F}(\varepsilon; \text{01}))\psi_\ell \left(\frac{\varepsilon}{2\mathrm{radius}(\mathcal{F}(\varepsilon; \text{01}))} \right), \tag{12.3}$$

which can sometimes lead to tighter results.

Some of our stronger results below will be stated for a restricted family of losses, originally explored by Bartlett, Jordan, and McAuliffe [2006]: namely, smooth losses whose convexity is quantified by a polynomial. Specifically, this restriction is characterized by the following condition.

Condition 12.3. *$\mathcal{F}$ is convex, with $\forall x \in \mathcal{X}, \sup_{f \in \mathcal{F}} |f(x)| \leq \bar{B}$ for some constant $\bar{B} \in (0,\infty)$, and there exists a pseudometric $d_\ell : [-\bar{B}, \bar{B}]^2 \to [0, \bar{d}_\ell]$ for some constant $\bar{d}_\ell \in (0,\infty)$, and constants $L, C_\ell \in (0,\infty)$ and $r_\ell \in (0,\infty]$ such that $\forall x, y \in [-\bar{B}, \bar{B}], |\ell(x) - \ell(y)| \leq Ld_\ell(x,y)$ and the function $\bar{\delta}_\ell(\varepsilon) = \inf \left\{ \frac{1}{2}\ell(x) + \frac{1}{2}\ell(y) - \ell(\frac{1}{2}x + \frac{1}{2}y) : x, y \in [-\bar{B}, \bar{B}], d_\ell(x,y) \geq \varepsilon \right\} \cup \{\infty\}$ satisfies $\forall \varepsilon \in [0,\infty), \bar{\delta}_\ell(\varepsilon) \geq C_\ell \varepsilon^{r_\ell}$.*

In particular, note that if $\mathcal{F}$ is convex, the functions in $\mathcal{F}$ are uniformly bounded, and ℓ is convex and continuous, Condition 12.3 is always satisfied (though possibly with $r_\ell = \infty$) by taking $d_\ell(x, y) = |x - y|/(4\bar{B})$.

12.2.2 A Few Examples of Loss Functions

Here we briefly mention a few loss functions ℓ in common practical use, all of which are classification-calibrated. These examples are taken directly from the work of Bartlett, Jordan, and McAuliffe [2006], which additionally discusses many other interesting examples of classification-calibrated loss functions and their corresponding ψ_ℓ functions.

Example 1 The *exponential loss* is specified as $\ell(x) = e^{-x}$. This loss function appears in many contexts in machine learning; for instance, the popular AdaBoost method can be viewed as an algorithm that greedily optimizes the exponential loss [Freund and Schapire, 1997]. Bartlett, Jordan, and McAuliffe [2006] show that under the exponential loss, $\psi_\ell(x) = 1 - \sqrt{1 - x^2}$, which is tightly approximated by $x^2/2$ for small x. They also show this loss satisfies the conditions on ℓ in Condition 12.3 with $d_\ell(x, y) = |x - y|$, $L = e^{\bar{B}}$, $C_\ell = e^{-\bar{B}}/8$, and $r_\ell = 2$.

Example 2 The *hinge loss*, specified as $\ell(x) = \max\{1 - x, 0\}$, is another common surrogate loss in machine learning practice today. For instance, it is used in the objective of the Support Vector Machine (along with a regularization term) [Cortes and Vapnik, 1995]. Bartlett, Jordan, and McAuliffe [2006] show that for the hinge loss, $\psi_\ell(x) = |x|$. The hinge loss is Lipschitz continuous, with Lipschitz constant 1. However, for the remaining conditions on ℓ in Condition 12.3, any $x, y \leq 1$ have $\frac{1}{2}\ell(x) + \frac{1}{2}\ell(y) = \ell(\frac{1}{2}x + \frac{1}{2}y)$, so that $\bar{\delta}_\ell(\varepsilon) = 0$; hence, $r_\ell = \infty$ is required.

Example 3 The *quadratic loss* (or squared loss), specified as $\ell(x) = (1 - x)^2$, is often used in so-called *plug-in* classifiers [Audibert and Tsybakov, 2007], which approach the problem of learning a classifier by estimating the regression function $\mathbb{E}[Y|X = x] = 2\eta(x) - 1$, and then

taking the sign of this estimator to get a binary classifier. The quadratic loss has the convenient property that for any distribution P over $\mathcal{X} \times \mathcal{Y}$, $h^*{}_P(\cdot) = 2\eta(\cdot; P) - 1$, so that it is straightforward to describe the set of distributions P satisfying the assumption $h^*{}_P \in \mathcal{F}$. Bartlett, Jordan, and McAuliffe [2006] show that for the quadratic loss, $\psi_\ell(x) = x^2$. They also show the quadratic loss satisfies the conditions on ℓ in Condition 12.3, with $L = 2(\bar{B} + 1)$, $C_\ell = 1/4$, and $r_\ell = 2$. In fact, they study the general family of losses $\ell(x) = |1 - x|^p$, for $p \in (1, \infty)$, and show that $\psi_\ell(x)$ and r_ℓ exhibit a range of behaviors varying with p.

Example 4 The *truncated quadratic loss* is specified as $\ell(x) = (\max\{1 - x, 0\})^2$. Bartlett, Jordan, and McAuliffe [2006] show that in this case, $\psi_\ell(x) = x^2$. They also show that, under the pseudometric $d_\ell(a, b) = |\min\{a, 1\} - \min\{b, 1\}|$, the truncated quadratic loss satisfies the conditions on ℓ in Condition 12.3, with $L = 2(\bar{B} + 1)$, $C_\ell = 1/4$, and $r_\ell = 2$.

12.2.3 Empirical ℓ-Risk Minimization

For any $m \in \mathbb{N}$, $g : \mathcal{X} \to \bar{\mathbb{R}}$, and $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$, define the *empirical ℓ-risk* as $\mathrm{R}_\ell(g; S) = m^{-1} \sum_{i=1}^m \ell(g(x_i) y_i)$. At times it will be convenient to keep track of the indices for a subsequence of $\mathcal{Z}$, and for this reason we also overload the notation, so that for any $Q = \{(i_1, y_1), \ldots, (i_m, y_m)\} \in (\mathbb{N} \times \mathcal{Y})^m$, we define $S[Q] = \{(X_{i_1}, y_1), \ldots, (X_{i_m}, y_m)\}$ and $\mathrm{R}_\ell(g; Q) = \mathrm{R}_\ell(g; S[Q])$. For completeness, we also generally define $\mathrm{R}_\ell(g; \emptyset) = 0$. The method of empirical ℓ-risk minimization, here denoted by $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$, is characterized by the property that it returns $\hat{h} = \mathrm{argmin}_{h \in \mathcal{H}} \mathrm{R}_\ell(h; \mathcal{Z}_m)$. This is a well-studied and classical passive learning method, presently in popular use in applications, and as such it will serve as our baseline for passive learning methods.

12.2.4 Localized Sample Complexities

The derivation of localized excess risk bounds can essentially be motivated as follows. Suppose we are interested in bounding the excess ℓ-risk of $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$. Further suppose we have a coarse guarantee $U_\ell(\mathcal{H}, m)$ on the excess ℓ-risk of the $\hat{h}$ returned by $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$: that is, $\mathrm{R}_\ell(\hat{h}) - \mathrm{R}_\ell(h^*) \le U_\ell(\mathcal{H}, m)$. In some sense, this guarantee identifies a set $\mathcal{H}' \subseteq \mathcal{H}$ of functions that a priori have the *potential* to be returned by $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$ (namely, $\mathcal{H}' = \mathcal{H}(U_\ell(\mathcal{H}, m); \ell)$), while those in $\mathcal{H} \setminus \mathcal{H}'$ do not. With this information in hand, we can think of $\mathcal{H}'$ as a kind of *effective* function class, and we can then think of $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$ as equivalent to $\mathrm{ERM}_\ell(\mathcal{H}', \mathcal{Z}_m)$. We may then repeat this same reasoning for $\mathrm{ERM}_\ell(\mathcal{H}', \mathcal{Z}_m)$, calculating $U_\ell(\mathcal{H}', m)$ to determine a set $\mathcal{H}'' = \mathcal{H}'(U_\ell(\mathcal{H}', m); \ell) \subseteq \mathcal{H}'$ of potential return values for *this* empirical minimizer, so that $\mathrm{ERM}_\ell(\mathcal{H}', \mathcal{Z}_m) = \mathrm{ERM}_\ell(\mathcal{H}'', \mathcal{Z}_m)$, and so on. This repeats until we identify a fixed-point set $\mathcal{H}^{(\infty)}$ of functions such that $\mathcal{H}^{(\infty)}(U_\ell(\mathcal{H}^{(\infty)}, m); \ell) = \mathcal{H}^{(\infty)}$, so that no further reduction is possible. Following this chain of reasoning back to the beginning, we find that $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m) = \mathrm{ERM}_\ell(\mathcal{H}^{(\infty)}, \mathcal{Z}_m)$, so that the function $\hat{h}$ returned by $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$ has excess ℓ-risk at most $U_\ell(\mathcal{H}^{(\infty)}, m)$, which may be significantly smaller than $U_\ell(\mathcal{H}, m)$, depending on how refined the original $U_\ell(\mathcal{H}, m)$ bound was.

To formalize this fixed-point argument for $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$, Koltchinskii [2006] makes use of the following quantities to define the coarse bound $U_\ell(\mathcal{H}, m)$ [see also Bartlett, Bousquet, and Mendelson, 2005, Giné and Koltchinskii, 2006]. For any $\mathcal{H} \subseteq [\mathcal{F}]$, $m \in \mathbb{N}$, $s \in [1, \infty)$, and any distribution P on $\mathcal{X} \times \mathcal{Y}$, letting $Q \sim P^m$, define

$$\phi_\ell(\mathcal{H}; m, P) = \mathbb{E}\left[\sup_{h,g \in \mathcal{H}} (\mathrm{R}_\ell(h; P) - \mathrm{R}_\ell(g; P)) - (\mathrm{R}_\ell(h; Q) - \mathrm{R}_\ell(g; Q)) \right],$$

$$\bar{U}_\ell(\mathcal{H}; P, m, s) = \bar{K}_1 \phi_\ell(\mathcal{H}; m, P) + \bar{K}_2 \mathrm{D}_\ell(\mathcal{H}; P)\sqrt{\frac{s}{m}} + \frac{\bar{K}_3 \bar{\ell} s}{m},$$

$$\tilde{U}_\ell(\mathcal{H}; P, m, s) = \tilde{K}\left(\phi_\ell(\mathcal{H}; m, P) + \mathrm{D}_\ell(\mathcal{H}; P)\sqrt{\frac{s}{m}} + \frac{\bar{\ell} s}{m} \right),$$

where $\bar{K}_1$, $\bar{K}_2$, $\bar{K}_3$, and $\tilde{K}$ are appropriately chosen constants.

We will be interested in having access to these quantities in the context of our algorithms;

however, since $\mathcal{P}_{XY}$ is not directly accessible to the algorithm, we will need to approximate these by data-dependent estimators. Toward this end, we define the following quantities, again taken from the work of Koltchinskii [2006]. For $\varepsilon > 0$, let $\mathbb{Z}_\varepsilon = \{j \in \mathbb{Z} : 2^j \geq \varepsilon\}$. For any $\mathcal{H} \subseteq [\mathcal{F}]$, $q \in \mathbb{N}$, and $S = \{(x_1, y_1), \ldots, (x_q, y_q)\} \in (\mathcal{X} \times \{-1, +1\})^q$, let $\mathcal{H}(\varepsilon; \ell, S) = \{h \in \mathcal{H} : \mathrm{R}_\ell(h; S) - \inf_{g \in \mathcal{H}} \mathrm{R}_\ell(g; S) \leq \varepsilon\}$; then for any sequence $\Xi = \{\xi_k\}_{k=1}^q \in \{-1, +1\}^q$, and any $s \in [1, \infty)$, define

$$\hat{\phi}_\ell(\mathcal{H}; S, \Xi) = \sup_{h,g \in \mathcal{H}} \frac{1}{q} \sum_{k=1}^q \xi_k \cdot \left(\ell(h(x_k)y_k) - \ell(g(x_k)y_k) \right),$$

$$\hat{\mathrm{D}}_\ell(\mathcal{H}; S)^2 = \sup_{h,g \in \mathcal{H}} \frac{1}{q} \sum_{k=1}^q \left(\ell(h(x_k)y_k) - \ell(g(x_k)y_k) \right)^2,$$

$$\hat{U}_\ell(\mathcal{H}; S, \Xi, s) = 12\hat{\phi}_\ell(\mathcal{H}; S, \Xi) + 34\hat{\mathrm{D}}_\ell(\mathcal{H}; S)\sqrt{\frac{s}{q}} + \frac{752\bar{\ell}s}{q}.$$

For completeness, define $\hat{\phi}_\ell(\mathcal{H}; \emptyset, \emptyset) = \hat{\mathrm{D}}_\ell(\mathcal{H}; \emptyset) = 0$, and $\hat{U}_\ell(\mathcal{H}; \emptyset, \emptyset, s) = 752\bar{\ell}s$.

The above quantities (with appropriate choices of $\bar{K}_1$, $\bar{K}_2$, $\bar{K}_3$, and $\tilde{K}$) can be formally related to each other and to the excess ℓ-risk of functions in $\mathcal{H}$ via the following general result; this variant is due to Koltchinskii [2006].

Lemma 12.4. *For any $\mathcal{H} \subseteq [\mathcal{F}]$, $s \in [1, \infty)$, distribution P over $\mathcal{X} \times \mathcal{Y}$, and any $m \in \mathbb{N}$, if $Q \sim P^m$ and $\Xi = \{\xi_1, \ldots, \xi_m\} \sim \mathrm{Uniform}(\{-1, +1\})^m$ are independent, and $h^* \in \mathcal{H}$ has $\mathrm{R}_\ell(h^*; P) = \inf_{h \in \mathcal{H}} \mathrm{R}_\ell(h; P)$, then with probability at least $1 - 6e^{-s}$, the following claims hold.*

$$\forall h \in \mathcal{H}, \mathrm{R}_\ell(h; P) - \mathrm{R}_\ell(h^*; P) \leq \mathrm{R}_\ell(h; Q) - \mathrm{R}_\ell(h^*; Q) + \bar{U}_\ell(\mathcal{H}; P, m, s),$$

$$\forall h \in \mathcal{H}, \mathrm{R}_\ell(h; Q) - \inf_{g \in \mathcal{H}} \mathrm{R}_\ell(g; Q) \leq \mathrm{R}_\ell(h; P) - \mathrm{R}_\ell(h^*; P) + \bar{U}_\ell(\mathcal{H}; P, m, s),$$

$$\bar{U}_\ell(\mathcal{H}; P, m, s) < \hat{U}_\ell(\mathcal{H}; Q, \Xi, s) < \tilde{U}_\ell(\mathcal{H}; P, m, s).$$

We typically expect the $\bar{U}$, $\hat{U}$, and $\tilde{U}$ quantities to be roughly within constant factors of each other. Following Koltchinskii [2006] and Giné and Koltchinskii [2006], we can use this result to derive localized bounds on the number of samples sufficient for $\mathrm{ERM}_\ell(\mathcal{H}, \mathcal{Z}_m)$ to achieve a given excess ℓ-risk. Specifically, for $\mathcal{H} \subseteq [\mathcal{F}]$, distribution P over $\mathcal{X} \times \mathcal{Y}$, values $\gamma, \gamma_1, \gamma_2 \geq 0$,

$s \in [1, \infty)$, and any function $\mathfrak{s} : (0, \infty)^2 \to [1, \infty)$, define the following quantities.

$$\bar{\mathrm{M}}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P, s) = \min \left\{ m \in \mathbb{N} : \bar{U}_\ell(\mathcal{H}(\gamma_2; \ell, P); P, m, s) < \gamma_1 \right\},$$

$$\bar{\mathrm{M}}_\ell(\gamma; \mathcal{H}, P, \mathfrak{s}) = \sup_{\gamma' \geq \gamma} \bar{\mathrm{M}}_\ell(\gamma'/2, \gamma'; \mathcal{H}, P, \mathfrak{s}(\gamma, \gamma')),$$

$$\tilde{\mathrm{M}}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P, s) = \min \left\{ m \in \mathbb{N} : \tilde{U}_\ell(\mathcal{H}(\gamma_2; \ell, P); P, m, s) \leq \gamma_1 \right\},$$

$$\tilde{\mathrm{M}}_\ell(\gamma; \mathcal{H}, P, \mathfrak{s}) = \sup_{\gamma' \geq \gamma} \tilde{\mathrm{M}}_\ell(\gamma'/2, \gamma'; \mathcal{H}, P, \mathfrak{s}(\gamma, \gamma')).$$

These quantities are well-defined for $\gamma_1, \gamma_2, \gamma > 0$ when $\lim_{m \to \infty} \phi_\ell(\mathcal{H}; m, P) = 0$. In other cases, for completeness, we define them to be ∞.

In particular, the quantity $\bar{\mathrm{M}}_\ell(\gamma; \mathcal{F}, \mathcal{P}_{XY}, \mathfrak{s})$ is used in Theorem 12.6 below to quantify the performance of $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$. The primary practical challenge in calculating $\bar{\mathrm{M}}_\ell(\gamma; \mathcal{H}, P, \mathfrak{s})$ is handling the $\phi_\ell(\mathcal{H}(\gamma'; \ell, P); m, P)$ quantity. In the literature, the typical (only?) way such calculations are approached is by first deriving a bound on $\phi_\ell(\mathcal{H}'; m, P)$ for every $\mathcal{H}' \subseteq \mathcal{H}$ in terms of some natural measure of complexity for the full class $\mathcal{H}$ (e.g., entropy numbers) and some very basic measure of complexity for $\mathcal{H}'$: most often $\mathrm{D}_\ell(\mathcal{H}'; P)$ and sometimes a seminorm of an envelope function for $\mathcal{H}'$. After this, one then proceeds to bound these basic measures of complexity for the specific subsets $\mathcal{H}(\gamma'; \ell, P)$, as a function of γ'. Composing these two results is then sufficient to bound $\phi_\ell(\mathcal{H}(\gamma'; \ell, P); m, P)$. For instance, bounds based on an entropy integral tend to follow this strategy. This approach effectively decomposes the problem of calculating the complexity of $\mathcal{H}(\gamma'; \ell, P)$ into the problem of calculating the complexity of $\mathcal{H}$ and the problem of calculating some much more basic properties of $\mathcal{H}(\gamma'; \ell, P)$. See [Bartlett, Jordan, and McAuliffe, 2006, Giné and Koltchinskii, 2006, Koltchinskii, 2006, van der Vaart and Wellner, 1996], or Section 12.5 below, for several explicit examples of this technique.

Another technique often (though not always) used in conjunction with the above strategy when deriving explicit rates of convergence is to relax $\mathrm{D}_\ell(\mathcal{H}(\gamma'; \ell, P); P)$ to $\mathrm{D}_\ell(\mathcal{F}^*(\gamma'; \ell, P); P)$ or $\mathrm{D}_\ell([\mathcal{H}](\gamma'; \ell, P); P)$. This relaxation can sometimes be a source of slack; however, in many interesting cases, such as for certain losses ℓ [e.g., Bartlett, Jordan, and McAuliffe, 2006], or

even certain noise conditions [e.g., Mammen and Tsybakov, 1999, Tsybakov, 2004], this relaxed quantity can still lead to nearly tight bounds.

For our purposes, it will be convenient to make these common techniques explicit in the results. In later sections, this will make the benefits of our proposed methods more explicit, while still allowing us to state results in a form abstract enough to capture the variety of specific complexity measures most often used in conjunction with the above approach. Toward this end, we have the following definition.

Definition 12.5. *For every distribution P over $\mathcal{X} \times \mathcal{Y}$, let $\mathring{\phi}_\ell(\sigma, \mathcal{H}; m, P)$ be a quantity defined for every $\sigma \in [0, \infty]$, $\mathcal{H} \subseteq [\mathcal{F}]$, and $m \in \mathbb{N}$, such that the following conditions are satisfied when $h^*_P \in \mathcal{H}$.*

$$\text{If } 0 \le \sigma \le \sigma', \mathcal{H} \subseteq \mathcal{H}' \subseteq [\mathcal{F}], \mathcal{U} \subseteq \mathcal{X}, \text{ and } m' \le m,$$

$$\text{then } \mathring{\phi}_\ell(\sigma, \mathcal{H}_{\mathcal{U}, h^*_P}; m, P) \le \mathring{\phi}_\ell(\sigma', \mathcal{H}'; m', P). \tag{12.4}$$

$$\forall \sigma \ge \mathrm{D}_\ell(\mathcal{H}; P), \phi_\ell(\mathcal{H}; m, P) \le \mathring{\phi}_\ell(\sigma, \mathcal{H}; m, P). \tag{12.5}$$

For instance, most bounds based on entropy integrals can be made to satisfy this. See Section 12.5.3 for explicit examples of quantities $\mathring{\phi}_\ell$ from the literature that satisfy this definition. Given a function $\mathring{\phi}_\ell$ of this type, we define the following quantity for $m \in \mathbb{N}$, $s \in [1, \infty)$, $\zeta \in [0, \infty]$, $\mathcal{H} \subseteq [\mathcal{F}]$, and a distribution P over $\mathcal{X} \times \mathcal{Y}$.

$$\mathring{U}_\ell(\mathcal{H}, \zeta; P, m, s)$$
$$= \tilde{K} \left(\mathring{\phi}_\ell(\mathrm{D}_\ell([\mathcal{H}](\zeta; \ell, P); P), \mathcal{H}; m, P) + \mathrm{D}_\ell([\mathcal{H}](\zeta; \ell, P); P) \sqrt{\frac{s}{m}} + \frac{\bar{\ell}s}{m} \right).$$

Note that when $h^*_P \in \mathcal{H}$, since $\mathrm{D}_\ell([\mathcal{H}](\gamma; \ell, P); P) \ge \mathrm{D}_\ell(\mathcal{H}(\gamma; \ell, P); P)$, Definition 12.5 implies $\phi_\ell(\mathcal{H}(\gamma; \ell, P); m, P) \le \mathring{\phi}_\ell(\mathrm{D}_\ell([\mathcal{H}](\gamma; \ell, P); P), \mathcal{H}(\gamma; \ell, P); P, m)$, and furthermore $\mathcal{H}(\gamma; \ell, P) \subseteq \mathcal{H}$ so that $\mathring{\phi}_\ell(\mathrm{D}_\ell([\mathcal{H}](\gamma; \ell, P); P), \mathcal{H}(\gamma; \ell, P); P, m) \le \mathring{\phi}_\ell(\mathrm{D}_\ell([\mathcal{H}](\gamma; \ell, P); P), \mathcal{H}; P, m)$. Thus,

$$\tilde{U}_\ell(\mathcal{H}(\gamma; \ell, P); P, m, s) \le \mathring{U}_\ell(\mathcal{H}(\gamma; \ell, P), \gamma; P, m, s) \le \mathring{U}_\ell(\mathcal{H}, \gamma; P, m, s). \tag{12.6}$$

233

Furthermore, when $h^*{}_P \in \mathcal{H}$, for any measurable $\mathcal{U} \subseteq \mathcal{U}' \subseteq \mathcal{X}$, any $\gamma' \geq \gamma \geq 0$, and any $\mathcal{H}' \subseteq [\mathcal{F}]$ with $\mathcal{H} \subseteq \mathcal{H}'$,

$$\mathring{U}_\ell(\mathcal{H}_{\mathcal{U}, h^*{}_P}, \gamma; P, m, s) \leq \mathring{U}_\ell(\mathcal{H}'_{\mathcal{U}', h^*{}_P}, \gamma'; P, m, s). \tag{12.7}$$

Note that the fact that we use $D_\ell([\mathcal{H}](\gamma; \ell, P); P)$ instead of $D_\ell(\mathcal{H}(\gamma; \ell, P); P)$ in the definition of $\mathring{U}_\ell$ is crucial for these inequalities to hold; specifically, it is not necessarily true that $D_\ell(\mathcal{H}_{\mathcal{U}, h^*{}_P}(\gamma; \ell, P); P) \leq D_\ell(\mathcal{H}_{\mathcal{U}', h^*{}_P}(\gamma; \ell, P); P)$, but it is always the case that $[\mathcal{H}_{\mathcal{U}, h^*{}_P}](\gamma; \ell, P) \subseteq [\mathcal{H}_{\mathcal{U}', h^*{}_P}](\gamma; \ell, P)$ when $h^*{}_P \in [\mathcal{H}]$, so that $D_\ell([\mathcal{H}_{\mathcal{U}, h^*{}_P}](\gamma; \ell, P); P) \leq D_\ell([\mathcal{H}_{\mathcal{U}', h^*{}_P}](\gamma; \ell, P); P)$.

Finally, for $\mathcal{H} \subseteq [\mathcal{F}]$, distribution P over $\mathcal{X} \times \mathcal{Y}$, values $\gamma, \gamma_1, \gamma_2 \geq 0$, $s \in [1, \infty)$, and any function $\mathfrak{s} : (0, \infty)^2 \to [1, \infty)$, define

$$\mathring{M}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P, s) = \min\left\{m \in \mathbb{N} : \mathring{U}_\ell(\mathcal{H}, \gamma_2; P, m, s) \leq \gamma_1\right\},$$

$$\mathring{M}_\ell(\gamma; \mathcal{H}, P, \mathfrak{s}) = \sup_{\gamma' \geq \gamma} \mathring{M}_\ell(\gamma'/2, \gamma'; \mathcal{H}, P, \mathfrak{s}(\gamma, \gamma')).$$

For completeness, define $\mathring{M}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P, s) = \infty$ when $\mathring{U}_\ell(\mathcal{H}, \gamma_2; P, m, s) > \gamma_1$ for every $m \in \mathbb{N}$.

It will often be convenient to isolate the terms in $\mathring{U}_\ell$ when inverting for a sufficient m, thus arriving at an upper bound on $\mathring{M}_\ell$. Specifically, define

$$\dot{M}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P, s) = \min\left\{m \in \mathbb{N} : D_\ell([\mathcal{H}](\gamma_2; \ell, P); P)\sqrt{\frac{s}{m}} + \frac{\bar{\ell}s}{m} \leq \gamma_1\right\},$$

$$\ddot{M}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P) = \min\left\{m \in \mathbb{N} : \mathring{\phi}_\ell\left(D_\ell([\mathcal{H}](\gamma_2; \ell, P); P), \mathcal{H}; P, m\right) \leq \gamma_1\right\}.$$

This way, for $\tilde{c} = 1/(2\tilde{K})$, we have

$$\mathring{M}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P, s) \leq \max\left\{\ddot{M}_\ell(\tilde{c}\gamma_1, \gamma_2; \mathcal{H}, P), \dot{M}_\ell(\tilde{c}\gamma_1, \gamma_2; \mathcal{H}, P, s)\right\}. \tag{12.8}$$

Also note that we clearly have

$$\dot{M}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P, s) \leq s \cdot \max\left\{\frac{4D_\ell([\mathcal{H}](\gamma_2; \ell, P); \ell, P)^2}{\gamma_1^2}, \frac{2\bar{\ell}}{\gamma_1}\right\}, \tag{12.9}$$

so that, in the task of bounding $\mathring{M}_\ell$, we can simply focus on bounding $\ddot{M}_\ell$.

We will express our main abstract results below in terms of the incremental values $\mathring{\mathrm{M}}_\ell(\gamma_1, \gamma_2; \mathcal{H}, \mathcal{P}_{XY}, s)$; the quantity $\mathring{\mathrm{M}}_\ell(\gamma; \mathcal{H}, \mathcal{P}_{XY}, \mathfrak{s})$ will also be useful in deriving analogous results for ERM_ℓ. When $h^*_P \in \mathcal{H}$, (12.6) implies

$$\bar{\mathrm{M}}_\ell(\gamma; \mathcal{H}, P, \mathfrak{s}) \leq \tilde{\mathrm{M}}_\ell(\gamma; \mathcal{H}, P, \mathfrak{s}) \leq \mathring{\mathrm{M}}_\ell(\gamma; \mathcal{H}, P, \mathfrak{s}). \tag{12.10}$$

12.3 Methods Based on Optimizing the Surrogate Risk

Perhaps the simplest way to make use of a surrogate loss function is to try to optimize $\mathrm{R}_\ell(h)$ over $h \in \mathcal{F}$, until identifying $h \in \mathcal{F}$ with $\mathrm{R}_\ell(h) - \mathrm{R}_\ell(h^*) < \Gamma_\ell(\varepsilon)$, at which point we are guaranteed $\mathrm{er}(h) - \mathrm{er}(h^*) \leq \varepsilon$. In this section, we briefly discuss some known results for this basic idea, along with a comment on the potential drawbacks of this approach for active learning.

12.3.1 Passive Learning: Empirical Risk Minimization

In the context of passive learning, the method of *empirical ℓ-risk minimization* is one of the most-studied methods for optimizing $\mathrm{R}_\ell(h)$ over $h \in \mathcal{F}$. Based on Lemma 12.4 and the above definitions, one can derive a bound on the number of labeled data points m sufficient for $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ to achieve a given excess error rate. Specifically, the following theorem is due to Koltchinskii [2006] (slightly modified here, following Giné and Koltchinskii [2006], to allow for general $\mathfrak{s}$ functions). It will serve as our baseline for comparison in the applications below.

Theorem 12.6. *Fix any function $\mathfrak{s} : (0, \infty)^2 \to [1, \infty)$. If $h^* \in \mathcal{F}$, then for any $m \geq \bar{\mathrm{M}}_\ell(\Gamma_\ell(\varepsilon); \mathcal{F}, \mathcal{P}_{XY}, \mathfrak{s})$, with probability at least $1 - \sum_{j \in \mathbb{Z}_{\Gamma_\ell(\varepsilon)}} 6e^{-\mathfrak{s}(\Gamma_\ell(\varepsilon), 2^j)}$, $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ produces a function $\hat{h}$ such that $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$.*

12.3.2 Negative Results for Active Learning

As mentioned, there are several active learning methods designed to optimize a general loss function [Beygelzimer, Dasgupta, and Langford, 2009, Koltchinskii, 2010]. However, it turns

out that for many interesting loss functions, the number of labels required for active learning to achieve a given excess surrogate risk value is not significantly smaller than that sufficient for passive learning by ERM_ℓ.

Specifically, consider a problem with $\mathcal{X} = \{x_0, x_1\}$, let $z \in (0, 1/2)$ be a constant, and for $\varepsilon \in (0, z)$, let $\mathcal{P}(\{x_1\}) = \varepsilon/(2z)$, $\mathcal{P}(\{x_0\}) = 1 - \mathcal{P}(\{x_1\})$, and suppose $\mathcal{F}$ and ℓ are such that for $\eta(x_1) = 1/2 + z$ and any $\eta(x_0) \in [4/6, 5/6]$, we have $h^* \in \mathcal{F}$. For this problem, any function h with $\mathrm{sign}(h(x_1)) \neq +1$ has $\mathrm{er}(h) - \mathrm{er}(h^*) \geq \varepsilon$, so that $\Gamma_\ell(\varepsilon) \leq (\varepsilon/(2z))(\ell^\star_-(\eta(x_1)) - \ell^\star(\eta(x_1)))$; when ℓ is classification-calibrated and $\bar{\ell} < \infty$, this is $c\varepsilon$, for some ℓ-dependent $c \in (0, \infty)$. Any function h with $\mathrm{R}_\ell(h) - \mathrm{R}_\ell(h^*) \leq c\varepsilon$ for this problem must have $\mathrm{R}_\ell(h; \mathcal{P}_{\{x_0\}}) - \mathrm{R}_\ell(h^*; \mathcal{P}_{\{x_0\}}) \leq c\varepsilon/\mathcal{P}(\{x_0\}) = O(\varepsilon)$. Existing results of Hanneke and Yang [2010] (with a slight modification to rescale for $\eta(x_0) \in [4/6, 5/6]$) imply that, for many classification-calibrated losses ℓ, the minimax optimal number of labels sufficient for an active learning algorithm to achieve this is $\Theta(1/\varepsilon)$. Hanneke and Yang [2010] specifically show this for losses ℓ that are strictly positive, decreasing, strictly convex, and twice differentiable with continuous second derivative; however, that result can easily be extended to a wide variety of other classification-calibrated losses, such as the quadratic loss, which satisfy these conditions in a neighborhood of 0. It is also known [Bartlett, Jordan, and McAuliffe, 2006] (see also below) that for many such losses (specifically, those satisfying Condition 12.3 with $r_\ell = 2$), $\Theta(1/\varepsilon)$ random labeled samples are sufficient for ERM_ℓ to achieve this same guarantee, so that results that only bound the surrogate risk of the function produced by an active learning method in this scenario can be at most a constant factor smaller than those provable for passive learning methods.

In the next section, we provide an active learning algorithm and a general analysis of its performance which, in the special case described above, guarantees excess error rate less than ε with high probability, using a number of label requests $O(\log(1/\varepsilon) \log \log(1/\varepsilon))$. The implication is that, to identify the improvements achievable by active learning with a surrogate loss, it is not sufficient to merely analyze the surrogate risk of the function produced by a given active learning

algorithm. Indeed, since we are not particularly interested in the surrogate risk itself, we may even consider active learning algorithms that do not actually optimize $R_\ell(h)$ over $h \in \mathcal{F}$ (even in the limit).

12.4   Alternative Use of the Surrogate Loss

Given that we are interested in ℓ only insofar as it helps us to optimize the error rate with computational efficiency, we should ask whether there is a method that sometimes makes more effective use of ℓ in terms of optimizing the error rate, while maintaining essentially the same computational advantages. The following method is essentially a relaxation of the methods of Koltchinskii [2010] and Hanneke [2012]. Similar results should also hold for analogous relaxations of the related methods of Balcan, Beygelzimer, and Langford [2006], Dasgupta, Hsu, and Monteleoni [2007a], Balcan, Beygelzimer, and Langford [2009], and Beygelzimer, Dasgupta, and Langford [2009].

Algorithm 1:

Input: surrogate loss ℓ, unlabeled sample budget u, labeled sample budget n

Output: classifier $\hat{h}$

---

0. $V \leftarrow \mathcal{F}, Q \leftarrow \{\}, m \leftarrow 1, t \leftarrow 0$

1. While $m < u$ and $t < n$

2. $\quad m \leftarrow m + 1$

3. $\quad$ If $X_m \in \mathrm{DIS}(V)$

4. $\quad\quad$ Request label Y_m and let $Q \leftarrow Q \cup \{(m, Y_m)\}, t \leftarrow t + 1$

5. $\quad$ If $\log_2(m) \in \mathbb{N}$

6. $\quad\quad V \leftarrow \left\{ h \in V : R_\ell(h; Q) - \inf_{g \in V} R_\ell(g; Q) \leq \hat{T}_\ell(V; Q, m) \right\}$

7. $\quad\quad Q \leftarrow \{\}$

8. Return $\hat{h} = \mathrm{argmin}_{h \in V} R_\ell(h; Q)$

The intuition behind this algorithm is that, since we are only interested in achieving low error rate, once we have identified $\text{sign}(h^*(x))$ for a given $x \in \mathcal{X}$, there is no need to further optimize the value $\mathbb{E}[\ell(\hat{h}(X)Y)|X = x]$. Thus, as long as we maintain $h^* \in V$, the data points $X_m \notin \text{DIS}(V)$ are typically less informative than those $X_m \in \text{DIS}(V)$. We therefore focus the label requests on those $X_m \in \text{DIS}(V)$, since there remains some uncertainty about $\text{sign}(h^*(X_m))$ for these points. The algorithm updates V periodically (Step 6), removing those functions h whose excess empirical risks (under the current sampling distribution) are relatively large; by setting this threshold $\hat{T}_\ell$ appropriately, we can guarantee the excess empirical risk of h^* is smaller than $\hat{T}_\ell$. Thus, the algorithm maintains $h^* \in V$ as an invariant, while focusing the sampling region $\text{DIS}(V)$.

In practice, the set V can be maintained implicitly, simply by keeping track of the constraints (Step 6) that define it; then the condition in Step 3 can be checked by solving two constraint satisfaction problems (one for each sign); likewise, the value $\inf_{g \in V} R_\ell(g; Q)$ in these constraints, as well as the final $\hat{h}$, can be found by solving constrained optimization problems. Thus, for convex loss functions and convex classes of function, these steps typically have computationally efficient realizations, as long as the $\hat{T}_\ell$ values can also be obtained efficiently. The quantity $\hat{T}_\ell$ in Algorithm 1 can be defined in one of several possible ways. In our present abstract context, we consider the following definition. Let $\{\xi'_k\}_{k \in \mathbb{N}}$ denote independent Rademacher random variables (i.e., uniform in $\{-1, +1\}$), also independent from $\mathcal{Z}$; these should be considered internal random bits used by the algorithm, which is therefore a randomized algorithm. For any $q \in \mathbb{N} \cup \{0\}$ and $Q = \{(i_1, y_1), \ldots, (i_q, y_q)\} \in (\mathbb{N} \times \{-1, +1\})^q$, let $S[Q] = \{(X_{i_1}, y_1), \ldots, (X_{i_q}, y_q)\}$, $\Xi[Q] = \{\xi'_{i_k}\}_{k=1}^q$. For $s \in [1, \infty)$, define

$$\hat{U}_\ell(\mathcal{H}; Q, s) = \hat{U}_\ell(\mathcal{H}; S[Q], \Xi[Q], s).$$

Then we can define the quantity $\hat{T}_\ell$ in the method above as

$$\hat{T}_\ell(\mathcal{H}; Q, m) = \hat{U}_\ell(\mathcal{H}; Q, \hat{\mathfrak{s}}(m)), \tag{12.11}$$

238

for some $\hat{\mathfrak{s}} : \mathbb{N} \to [1, \infty)$. This definition has the appealing property that it allows us to interpret the update in Step 6 in two complementary ways: as comparing the empirical risks of functions in V under the conditional distribution given the region of disagreement $\mathcal{P}_{\mathrm{DIS}(V)}$, and as comparing the empirical risks of the functions in $V_{\mathrm{DIS}(V)}$ under the original distribution $\mathcal{P}_{XY}$. Our abstract results below are based on this definition of $\hat{T}_\ell$. This can sometimes be problematic due to the computational challenge of the optimization problem in the definitions of $\hat{\phi}_\ell$ and $\hat{\mathrm{D}}_\ell$. There has been considerable work on calculating and bounding $\hat{\phi}_\ell$ for various classes $\mathcal{F}$ and losses ℓ [e.g., Bartlett and Mendelson, 2002, Koltchinskii, 2001], but it is not always feasible. However, the specific applications below continue to hold if we instead take $\hat{T}_\ell$ based on a well-chosen upper bound on the respective $\mathring{U}_\ell$ function, such as those obtained in the derivations of those respective results below; we provide descriptions of such efficiently-computable relaxations for each of the applications below (though in some cases, these bounds have a mild dependence on $\mathcal{P}_{XY}$ via certain parameters of the specific noise conditions considered there).

We have the following theorem, which represents our main abstract result. The proof is included in Appendix 12.6.

Theorem 12.7. *Fix any function* $\hat{\mathfrak{s}} : \mathbb{N} \to [1, \infty)$. *Let* $j_\ell = -\lceil \log_2(\bar{\ell}) \rceil$, *define* $u_{j_\ell - 2} = u_{j_\ell - 1} = 1$, *and for each integer* $j \geq j_\ell$, *let* $\mathcal{F}_j = \mathcal{F}(\mathcal{E}_\ell(2^{2-j}); {}_{01})_{\mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{2-j}); 01))}$, $\mathcal{U}_j = \mathrm{DIS}(\mathcal{F}_j)$, *and suppose* $u_j \in \mathbb{N}$ *satisfies* $\log_2(u_j) \in \mathbb{N}$ *and*

$$u_j \geq 2\mathring{\mathrm{M}}_\ell(2^{-j-1}, 2^{2-j}; \mathcal{F}_j, \mathcal{P}_{XY}, \hat{\mathfrak{s}}(u_j)) \vee u_{j-1} \vee 2u_{j-2}. \tag{12.12}$$

Suppose $h^* \in \mathcal{F}$. *For any* $\varepsilon \in (0, 1)$ *and* $s \in [1, \infty)$, *letting* $j_\varepsilon = \lceil \log_2(1/\Gamma_\ell(\varepsilon)) \rceil$, *if*

$$u \geq u_{j_\varepsilon} \qquad \text{and} \qquad n \geq s + 2e \sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j) u_j,$$

then, with arguments ℓ, u, *and* n, *Algorithm 1 uses at most* u *unlabeled samples and makes at most* n *label requests, and with probability at least*

$$1 - 2^{-s} - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)},$$

239

returns a function $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$.

The number of label requests indicated by Theorem 12.7 can often (though not always) be significantly smaller than the number of random labeled data points sufficient for ERM_ℓ to achieve the same, as indicated by Theorem 12.6. This is typically the case when $\mathcal{P}(\mathcal{U}_j) \to 0$ as $j \to \infty$. When this is the case, the number of labels requested by the algorithm is sublinear in the number of unlabeled samples it processes; below, we will derive more explicit results for certain types of function classes $\mathcal{F}$, by characterizing the rate at which $\mathcal{P}(\mathcal{U}_j)$ vanishes in terms of a complexity measure known as the disagreement coefficient.

For the purpose of calculating the values $\mathring{\mathrm{M}}_\ell$ in Theorem 12.7, it is sometimes convenient to use the alternative interpretation of Algorithm 1, in terms of sampling Q from the conditional distribution $\mathcal{P}_{\mathrm{DIS}(V)}$. Specifically, the following lemma allows us to replace calculations in terms of $\mathcal{F}_j$ and $\mathcal{P}_{XY}$ with calculations in terms of $\mathcal{F}(\mathcal{E}_\ell(2^{1-j}); {}_{01})$ and $\mathcal{P}_{\mathrm{DIS}(\mathcal{F}_j)}$. Its proof is included in Appendix 12.6

Lemma 12.8. *Let $\mathring{\phi}_\ell$ be any function satisfying Definition 12.5. Let P be any distribution over $\mathcal{X} \times \mathcal{Y}$. For any measurable $\mathcal{U} \subseteq \mathcal{X} \times \mathcal{Y}$ with $P(\mathcal{U}) > 0$, define $P_\mathcal{U}(\cdot) = P(\cdot|\mathcal{U})$. Also, for any $\sigma \geq 0$, $\mathcal{H} \subseteq [\mathcal{F}]$, and $m \in \mathbb{N}$, if $P\left(\overline{\mathrm{DISF}}(\mathcal{H})\right) > 0$, define*

$$
\mathring{\phi}'_\ell(\sigma, \mathcal{H}; m, P) =
$$
$$
32 \left(\inf_{\substack{\mathcal{U}=\mathcal{U}'\times\mathcal{Y}: \\ \mathcal{U}'\supseteq\mathrm{DISF}(\mathcal{H})}} P(\mathcal{U})\mathring{\phi}_\ell\left(\frac{\sigma}{\sqrt{P(\mathcal{U})}}, \mathcal{H}; \lceil (1/2)P(\mathcal{U})m \rceil, P_\mathcal{U} \right) + \frac{\bar{\ell}}{m} + \sigma\sqrt{\frac{1}{m}} \right), \quad (12.13)
$$

and otherwise define $\mathring{\phi}'_\ell(\sigma, \mathcal{H}; m, P) = 0$. Then the function $\mathring{\phi}'_\ell$ also satisfies Definition 12.5.

Plugging this $\mathring{\phi}'_\ell$ function into Theorem 12.7 immediately yields the following corollary, the proof of which is included in Appendix 12.6.

Corollary 12.9. *Fix any function $\hat{\mathfrak{s}} : \mathbb{N} \to [1, \infty)$. Let $j_\ell = -\lceil \log_2(\bar{\ell}) \rceil$, define $u_{j_\ell-2} = u_{j_\ell-1} = 1$, and for each integer $j \geq j_\ell$, let $\mathcal{F}_j$ and $\mathcal{U}_j$ be as in Theorem 12.7, and if $\mathcal{P}(\mathcal{U}_j) > 0$, suppose*

$u_j \in \mathbb{N}$ *satisfies* $\log_2(u_j) \in \mathbb{N}$ *and*

$$u_j \geq 4\mathcal{P}(\mathcal{U}_j)^{-1}\mathring{\mathrm{M}}_\ell\left(\frac{2^{-j-7}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{2-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, \hat{\mathfrak{s}}(u_j)\right) \vee u_{j-1} \vee 2u_{j-2}. \tag{12.14}$$

If $\mathcal{P}(\mathcal{U}_j) = 0$, *let* $u_j \in \mathbb{N}$ *satisfy* $\log_2(u_j) \in \mathbb{N}$ *and* $u_j \geq \tilde{K}\bar{\ell}\hat{\mathfrak{s}}(u_j)2^{j+2} \vee u_j \vee 2u_{j-2}$. *Suppose* $h^* \in \mathcal{F}$. *For any* $\varepsilon \in (0,1)$ *and* $s \in [1, \infty)$, *letting* $j_\varepsilon = \lceil \log_2(1/\Gamma_\ell(\varepsilon))\rceil$, *if*

$$u \geq u_{j_\varepsilon} \qquad \qquad \text{and} \qquad \qquad n \geq s + 2e\sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j)u_j,$$

then, with arguments ℓ, u, *and* n, *Algorithm 1 uses at most* u *unlabeled samples and makes at most* n *label requests, and with probability at least*

$$1 - 2^{-s} - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)},$$

returns a function $\hat{h}$ *with* $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$.

Algorithm 1 can be modified in a variety of interesting ways, leading to related methods that can be analyzed analogously. One simple modification is to use a more involved bound to define the quantity $\hat{T}_\ell$. For instance, for Q as above, and a function $\hat{\mathfrak{s}} : (0, \infty) \times \mathbb{Z} \times \mathbb{N} \to [1, \infty)$, one could define

$$\hat{T}_\ell(\mathcal{H}; Q, m) = (3/2)q^{-1}\inf\left\{\lambda > 0 : \forall k \in \mathbb{Z}_\lambda, \right.$$
$$\left. \hat{U}_\ell\left(\mathcal{H}\left(3q^{-1}2^{k-1}; \ell, S[Q]\right); Q, \hat{\mathfrak{s}}(\lambda, k, m)\right) \leq 2^{k-4}q^{-1}\right\},$$

for which one can also prove a result similar to Lemma 12.4 [see Giné and Koltchinskii, 2006, Koltchinskii, 2006]. This definition shares the convenient dual-interpretations property mentioned above about $\hat{U}_\ell(\mathcal{H}; Q, \hat{\mathfrak{s}}(m))$; furthermore, results analogous to those above for Algorithm 1 also hold under this definition (under mild restrictions on the allowed $\hat{\mathfrak{s}}$ functions), with only a few modifications to constants and event probabilities (e.g., summing over the $k \in \mathbb{Z}_\lambda$ argument to $\hat{\mathfrak{s}}$ in the probability, while setting the λ argument to 2^{-j} for the largest j with $u_j \leq 2^i$).

The update trigger in Step 5 can also be modified in several ways, leading to interesting related methods. One possibility is that, if we have updated the V set $k-1$ times already, and

the previous update occurred at $m = m_{k-1}$, at which point $V = V_{k-1}$, $Q = Q_{k-1}$ (before the update), then we could choose to update V a kth time when $\log_2(m - m_{k-1}) \in \mathbb{N}$ and $\hat{U}_\ell(V; Q, \hat{\mathfrak{s}}(\hat{\gamma}_{k-1}, m - m_{k-1})) \frac{|Q| \vee 1}{m - m_{k-1}} \leq \hat{\gamma}_{k-1}/2$, for some function $\hat{\mathfrak{s}} : (0, \infty) \times \mathbb{N} \to [1, \infty)$, where $\hat{\gamma}_{k-1}$ is inductively defined as $\hat{\gamma}_{k-1} = \hat{U}_\ell(V_{k-1}; Q_{k-1}, \hat{\mathfrak{s}}(\hat{\gamma}_{k-2}, m_{k-1} - m_{k-2})) \frac{|Q_{k-1}| \vee 1}{m_{k-1} - m_{k-2}}$ (and $\hat{\gamma}_0 = \bar{\ell}$), and we would then use $\hat{U}_\ell(V; Q, \hat{\mathfrak{s}}(\hat{\gamma}_{k-1}, m - m_{k-1}))$ for the $\hat{T}_\ell$ value in the update; in other words, we could update V when the value of the concentration inequality used in the update has been reduced by a factor of 2. This modification leads to results quite similar to those stated above (under mild restrictions on the allowed $\hat{\mathfrak{s}}$ functions), with only a change to the probability (namely, summing the exponential failure probabilities $e^{-\hat{\mathfrak{s}}(2^{-j}, 2^i)}$ over values of j between j_ℓ and j_ε, and values of i between 1 and $\log_2(u_j)$); additionally, with this modification, because we check for $\log_2(m - m_{k-1}) \in \mathbb{N}$ rather than $\log_2(m) \in \mathbb{N}$, one can remove the "$\vee u_{j-1} \vee 2u_{j-2}$" term in (12.12) and (12.14) (though this has no effect for the applications below). Another interesting possibility in this vein is to update when $\log_2(m - m_{k-1}) \in \mathbb{N}$ and $\hat{U}_\ell(V; Q, \hat{\mathfrak{s}}(\Gamma_\ell(2^{-k}), m - m_{k-1})) \frac{|Q| \vee 1}{m - m_{k-1}} < \Gamma_\ell(2^{-k})$. Of course, the value $\Gamma_\ell(2^{-k})$ is typically not directly available to us, but we could substitute a distribution-independent lower bound on $\Gamma_\ell(2^{-k})$, for instance based on the ψ_ℓ function of Bartlett, Jordan, and McAuliffe [2006]; in the active learning context, we could potentially use unlabeled samples to estimate a $\mathcal{P}$-dependent lower bound on $\Gamma_\ell(2^{-k})$, or even $\text{diam}(V)\psi_\ell(2^{-k}/2\text{diam}(V))$, based on (12.3), where $\text{diam}(V) = \sup_{h,g \in V} \Delta(h, g)$.

12.5 Applications

In this section, we apply the abstract results from above to a few commonly-studied scenarios: namely, VC subgraph classes and entropy conditions, with some additional mention of VC major classes and VC hull classes. In the interest of making the results more concise and explicit, we express them in terms of well-known conditions relating distances to excess risks. We also express them in terms of a lower bound on $\Gamma_\ell(\varepsilon)$ of the type in (12.2), with convenient properties

that allow for closed-form expression of the results. To simplify the presentation, we often omit numerical constant factors in the inequalities below, and for this we use the common notation $f(x) \lesssim g(x)$ to mean that $f(x) \leq cg(x)$ for some implicit universal constant $c \in (0, \infty)$.

12.5.1 Diameter Conditions

To begin, we first state some general characterizations relating distances to excess risks; these characterizations will make it easier to express our results more concretely below, and make for a more straightforward comparison between results for the above methods. The following condition, introduced by Mammen and Tsybakov [1999] and Tsybakov [2004], is a well-known noise condition, about which there is now an extensive literature [e.g., Bartlett, Jordan, and McAuliffe, 2006, Hanneke, 2011, 2012, Koltchinskii, 2006].

Condition 12.10. *For some* $a \in [1, \infty)$ *and* $\alpha \in [0, 1]$*, for every* $g \in \mathcal{F}^*$,

$$\Delta\left(g, h^*\right) \leq a\left(\mathrm{er}(g) - \mathrm{er}(h^*)\right)^\alpha.$$

Condition 12.10 can be equivalently expressed in terms of certain noise conditions [Bartlett, Jordan, and McAuliffe, 2006, Mammen and Tsybakov, 1999, Tsybakov, 2004]. Specifically, satisfying Condition 12.10 with some $\alpha < 1$ is equivalent to the existence of some $a' \in [1, \infty)$ such that, for all $\varepsilon > 0$,

$$\mathcal{P}\left(x : |\eta(x) - 1/2| \leq \varepsilon\right) \leq a'\varepsilon^{\alpha/(1-\alpha)},$$

which is often referred to as a *low noise* condition. Additionally, satisfying Condition 12.10 with $\alpha = 1$ is equivalent to having some $a' \in [1, \infty)$ such that

$$\mathcal{P}\left(x : |\eta(x) - 1/2| \leq 1/a'\right) = 0,$$

often referred to as a *bounded noise* condition.

For simplicity, we formulate our results in terms of a and α from Condition 12.10. However, for the abstract results in this section, the results remain valid under the weaker condition that

replaces $\mathcal{F}^*$ by $\mathcal{F}$, and adds the condition that $h^* \in \mathcal{F}$. In fact, the specific results in this section also remain valid using this weaker condition while additionally using (12.3) in place of (12.2), as remarked above.

An analogous condition can be defined for the surrogate loss function, as follows. Similar notions have been explored by Bartlett, Jordan, and McAuliffe [2006] and Koltchinskii [2006].

Condition 12.11. *For some $b \in [1, \infty)$ and $\beta \in [0, 1]$, for every $g \in [\mathcal{F}]$,*

$$\mathrm{D}_\ell \left(g, h^*{}_P; P\right)^2 \leq b \left(\mathrm{R}_\ell(g; P) - \mathrm{R}_\ell(h^*{}_P; P)\right)^\beta.$$

Note that these conditions are *always* satisfied for *some* values of a, b, α, β, since $\alpha = \beta = 0$ trivially satisfies the conditions. However, in more benign scenarios, values of α and β strictly greater than 0 can be satisfied. Furthermore, for some loss functions ℓ, Condition 12.11 can even be satisfied *universally*, in the sense that a value of $\beta > 0$ is satisfied for *all* distributions. In particular, Bartlett, Jordan, and McAuliffe [2006] show that this is the case under Condition 12.3, as stated in the following lemma [see Bartlett, Jordan, and McAuliffe, 2006, for the proof].

Lemma 12.12. *Suppose Condition 12.3 is satisfied. Let $\beta = \min\{1, \frac{2}{r_\ell}\}$ and $b = (2C_\ell \bar{d}_\ell^{\min\{r_\ell - 2, 0\}})^{-\beta} L^2$. Then* every *distribution P over $\mathcal{X} \times \mathcal{Y}$ with $h^*{}_P \in [\mathcal{F}]$ satisfies Condition 12.11 with these values of b and β.*

Under Condition 12.10, it is particularly straightforward to obtain bounds on $\Gamma_\ell(\varepsilon)$ based on a function $\Psi_\ell(\varepsilon)$ satisfying (12.2). For instance, since $x \mapsto x\psi_\ell(1/x)$ is nonincreasing on $(0, \infty)$ [Bartlett, Jordan, and McAuliffe, 2006], the function

$$\Psi_\ell(\varepsilon) = a\varepsilon^\alpha \psi_\ell \left(\varepsilon^{1-\alpha}/(2a)\right) \tag{12.15}$$

satisfies $\Psi_\ell(\varepsilon) \leq \Gamma_\ell(\varepsilon)$ [Bartlett, Jordan, and McAuliffe, 2006]. Furthermore, for classification-calibrated ℓ, Ψ_ℓ in (12.15) is strictly increasing, nonnegative, and continuous on $(0, 1)$ [Bartlett, Jordan, and McAuliffe, 2006], and has $\Psi_\ell(0) = 0$; thus, the inverse $\Psi_\ell^{-1}(\gamma)$, defined for all $\gamma > 0$ by

$$\Psi_\ell^{-1}(\gamma) = \inf\{\varepsilon > 0 : \gamma \leq \Psi_\ell(\varepsilon)\} \cup \{1\}, \tag{12.16}$$

is strictly increasing, nonnegative, and continuous on $(0, \Psi_\ell(1))$. Furthermore, one can easily show $x \mapsto \Psi_\ell^{-1}(x)/x$ is nonincreasing on $(0, \infty)$. Also note that $\forall \gamma > 0, \mathcal{E}_\ell(\gamma) \leq \Psi_\ell^{-1}(\gamma)$.

12.5.2 The Disagreement Coefficient

In order to more concisely state our results, it will be convenient to bound $\mathcal{P}(\mathrm{DIS}(\mathcal{H}))$ by a linear function of $\mathrm{radius}(\mathcal{H})$, for $\mathrm{radius}(\mathcal{H})$ in a given range. This type of relaxation has been used extensively in the active learning literature [Balcan, Hanneke, and Vaughan, 2010, Beygelzimer, Dasgupta, and Langford, 2009, Dasgupta, Hsu, and Monteleoni, 2007a, Friedman, 2009, Hanneke, 2007a, 2009, 2011, 2012, Koltchinskii, 2010, Mahalanabis, 2011, Raginsky and Rakhlin, 2011, Wang, 2011], and the coefficient in the linear function is typically referred to as the *disagreement coefficient*. Specifically, the following definition is due to Hanneke [2007a, 2011]; related quantities have been explored by Alexander [1987] and Giné and Koltchinskii [2006].

Definition 12.13. *For any $r_0 > 0$, define the* disagreement coefficient *of a function $h : \mathcal{X} \to \mathbb{R}$ with respect to $\mathcal{F}$ under $\mathcal{P}$ as*

$$\theta_h(r_0) = \sup_{r > r_0} \frac{\mathcal{P}(\mathrm{DIS}(\mathrm{B}(h, r)))}{r} \vee 1.$$

If $h^* \in \mathcal{F}$, define the disagreement coefficient of the class $\mathcal{F}$ as $\theta(r_0) = \theta_{h^*}(r_0)$.

The value of $\theta(\varepsilon)$ has been studied and bounded for various function classes $\mathcal{F}$ under various conditions on $\mathcal{P}$. In many cases of interest, $\theta(\varepsilon)$ is known to be bounded by a finite constant [Balcan, Hanneke, and Vaughan, 2010, Friedman, 2009, Hanneke, 2007a, 2011, Mahalanabis, 2011], while in other cases, $\theta(\varepsilon)$ may have an interesting dependence on ε [Balcan, Hanneke, and Vaughan, 2010, Raginsky and Rakhlin, 2011, Wang, 2011]. The reader is referred to the works of Hanneke [2011, 2012] for detailed discussions on the disagreement coefficient.

12.5.3 Specification of $\mathring{\phi}_\ell$

Next, we recall a few well-known bounds on the ϕ_ℓ function, which leads to a more concrete instance of a function $\mathring{\phi}_\ell$ satisfying Definition 12.5. Below, we let $\mathcal{G}^*$ denote the set of measurable functions $g : \mathcal{X} \times \mathcal{Y} \to \bar{\mathbb{R}}$. Also, for $\mathcal{G} \subseteq \mathcal{G}^*$, let $\mathrm{F}(\mathcal{G}) = \sup_{g \in \mathcal{G}} |g|$ denote the minimal *envelope* function for $\mathcal{G}$, and for $g \in \mathcal{G}^*$ let $\|g\|_P^2 = \int g^2 \mathrm{d}P$ denote the squared $L_2(P)$ seminorm of g; we will generally assume $\mathrm{F}(\mathcal{G})$ is measurable in the discussion below.

Uniform Entropy: The first bound is based on the work of van der Vaart and Wellner [2011]; related bounds have been studied by Giné and Koltchinskii [2006], Giné, Koltchinskii, and Wellner [2003], van der Vaart and Wellner [1996], and others. For a distribution P over $\mathcal{X} \times \mathcal{Y}$, a set $\mathcal{G} \subseteq \mathcal{G}^*$, and $\varepsilon \geq 0$, let $\mathcal{N}(\varepsilon, \mathcal{G}, L_2(P))$ denote the size of a minimal ε-cover of $\mathcal{G}$ (that is, the minimum number of balls of radius at most ε sufficient to cover $\mathcal{G}$), where distances are measured in terms of the $L_2(P)$ pseudo-metric: $(f, g) \mapsto \|f - g\|_P$. For $\sigma \geq 0$ and $\mathrm{F} \in \mathcal{G}^*$, define the function

$$J(\sigma, \mathcal{G}, \mathrm{F}) = \sup_Q \int_0^\sigma \sqrt{1 + \ln \mathcal{N}(\varepsilon \|\mathrm{F}\|_Q, \mathcal{G}, L_2(Q))} \mathrm{d}\varepsilon,$$

where Q ranges over all finitely discrete probability measures.

Fix any distribution P over $\mathcal{X} \times \mathcal{Y}$ and any $\mathcal{H} \subseteq [\mathcal{F}]$ with $h^*{}_P \in \mathcal{H}$, and let

$$\mathcal{G}_\mathcal{H} = \{(x, y) \mapsto \ell(h(x)y) : h \in \mathcal{H}\},$$

$$\text{and } \mathcal{G}_{\mathcal{H},P} = \{(x, y) \mapsto \ell(h(x)y) - \ell(h^*{}_P(x)y) : h \in \mathcal{H}\}. \tag{12.17}$$

Then, since $J(\sigma, \mathcal{G}_\mathcal{H}, \mathrm{F}) = J(\sigma, \mathcal{G}_{\mathcal{H},P}, \mathrm{F})$, it follows from Theorem 2.1 of van der Vaart and Wellner [2011] (and a triangle inequality) that for some universal constant $c \in [1, \infty)$, for any $m \in \mathbb{N}$, $\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})$, and $\sigma \geq \mathrm{D}_\ell(\mathcal{H}; P)$,

$$\phi_\ell(\mathcal{H}; P, m) \leq \tag{12.18}$$

$$cJ\left(\frac{\sigma}{\|\mathrm{F}\|_P}, \mathcal{G}_\mathcal{H}, \mathrm{F}\right) \|\mathrm{F}\|_P \left(\frac{1}{\sqrt{m}} + \frac{J\left(\frac{\sigma}{\|\mathrm{F}\|_P}, \mathcal{G}_\mathcal{H}, \mathrm{F}\right) \|\mathrm{F}\|_P \bar{\ell}}{\sigma^2 m}\right).$$

Based on (12.18), it is straightforward to define a function $\mathring{\phi}_\ell$ that satisfies Definition 12.5. Specifically, define

$$\mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P) =$$

$$\inf_{\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})} \inf_{\lambda \geq \sigma} cJ\left(\frac{\lambda}{\|\mathrm{F}\|_P}, \mathcal{G}_\mathcal{H}, \mathrm{F}\right) \|\mathrm{F}\|_P \left(\frac{1}{\sqrt{m}} + \frac{J\left(\frac{\lambda}{\|\mathrm{F}\|_P}, \mathcal{G}_\mathcal{H}, \mathrm{F}\right) \|\mathrm{F}\|_P \bar{\ell}}{\lambda^2 m}\right), \quad (12.19)$$

for c as in (12.18). By (12.18), $\mathring{\phi}_\ell^{(1)}$ satisfies (12.5). Also note that $m \mapsto \mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P)$ is non-increasing, while $\sigma \mapsto \mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P)$ is nondecreasing. Furthermore, $\mathcal{H} \mapsto \mathcal{N}(\varepsilon, \mathcal{G}_\mathcal{H}, L_2(Q))$ is nondecreasing for all Q, so that $\mathcal{H} \mapsto J(\sigma, \mathcal{G}_\mathcal{H}, \mathrm{F})$ is nondecreasing as well; since $\mathcal{H} \mapsto \mathrm{F}(\mathcal{G}_{\mathcal{H},P})$ is also nondecreasing, we see that $\mathcal{H} \mapsto \mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P)$ is nondecreasing. Similarly, for $\mathcal{U} \subseteq \mathcal{X}$, $\mathcal{N}(\varepsilon, \mathcal{G}_{\mathcal{H}_{\mathcal{U},h^*P}}, L_2(Q)) \leq \mathcal{N}(\varepsilon, \mathcal{G}_\mathcal{H}, L_2(Q))$ for all Q, so that $J(\sigma, \mathcal{G}_{\mathcal{H}_{\mathcal{U},h^*P}}, \mathrm{F}) \leq J(\sigma, \mathcal{G}_\mathcal{H}, \mathrm{F})$; because $\mathrm{F}(\mathcal{G}_{\mathcal{H}_{\mathcal{U},h^*P},P}) \leq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})$, we have $\mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}_{\mathcal{U},h^*P}; m, P) \leq \mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P)$ as well. Thus, to satisfy Definition 12.5, it suffices to take $\mathring{\phi}_\ell = \mathring{\phi}_\ell^{(1)}$.

Bracketing Entropy: Our second bound is a classic result in empirical process theory. For functions $g_1 \leq g_2$, a *bracket* $[g_1, g_2]$ is the set of functions $g \in \mathcal{G}^*$ with $g_1 \leq g \leq g_2$; $[g_1, g_2]$ is called an ε-bracket under $L_2(P)$ if $\|g_1 - g_2\|_P < \varepsilon$. Then $\mathcal{N}_{[]}(\varepsilon, \mathcal{G}, L_2(P))$ denotes the smallest number of ε-brackets (under $L_2(P)$) sufficient to cover $\mathcal{G}$. For $\sigma \geq 0$, define the function

$$J_{[]}(\sigma, \mathcal{G}, P) = \int_0^\sigma \sqrt{1 + \ln \mathcal{N}_{[]}(\varepsilon, \mathcal{G}, L_2(P))} \, d\varepsilon.$$

Fix any $\mathcal{H} \subseteq [\mathcal{F}]$, and let $\mathcal{G}_\mathcal{H}$ and $\mathcal{G}_{\mathcal{H},P}$ be as above. Then since $J_{[]}(\sigma, \mathcal{G}_\mathcal{H}, P) = J_{[]}(\sigma, \mathcal{G}_{\mathcal{H},P}, P)$, Lemma 3.4.2 of van der Vaart and Wellner [1996] and a triangle inequality imply that for some universal constant $c \in [1, \infty)$, for any $m \in \mathbb{N}$ and $\sigma \geq \mathrm{D}_\ell(\mathcal{H}; P)$,

$$\phi_\ell(\mathcal{H}; P, m) \leq cJ_{[]}(\sigma, \mathcal{G}_\mathcal{H}, P) \left(\frac{1}{\sqrt{m}} + \frac{J_{[]}(\sigma, \mathcal{G}_\mathcal{H}, P) \bar{\ell}}{\sigma^2 m}\right). \quad (12.20)$$

As-is, the right side of (12.20) nearly satisfies Definition 12.5 already. Only a slight modification is required to fulfill the requirement of monotonicity in σ. Specifically, define

$$\mathring{\phi}_\ell^{(2)}(\sigma, \mathcal{H}; P, m) = \inf_{\lambda \geq \sigma} cJ_{[]}(\lambda, \mathcal{G}_\mathcal{H}, P) \left(\frac{1}{\sqrt{m}} + \frac{J_{[]}(\lambda, \mathcal{G}_\mathcal{H}, P) \bar{\ell}}{\lambda^2 m}\right), \quad (12.21)$$

for c as in (12.20). Then taking $\mathring{\phi}_\ell = \mathring{\phi}_\ell^{(2)}$ suffices to satisfy Definition 12.5.

Since Definition 12.5 is satisfied for both $\mathring{\phi}_\ell^{(1)}$ and $\mathring{\phi}_\ell^{(2)}$, it is also satisfied for

$$\mathring{\phi}_\ell = \min\left\{\mathring{\phi}_\ell^{(1)}, \mathring{\phi}_\ell^{(2)}\right\}. \tag{12.22}$$

For the remainder of this section, we suppose $\mathring{\phi}_\ell$ is defined as in (12.22) (for all distributions P over $\mathcal{X} \times \mathcal{Y}$), and study the implications arising from the combination of this definition with the abstract theorems above.

12.5.4 VC Subgraph Classes

For a collection $\mathcal{A}$ of sets, a set $\{z_1, \ldots, z_k\}$ of points is said to be *shattered* by $\mathcal{A}$ if $|\{A \cap \{z_1, \ldots, z_k\} : A \in \mathcal{A}\}| = 2^k$. The VC dimension $\mathrm{vc}(\mathcal{A})$ of $\mathcal{A}$ is then defined as the largest integer k for which there exist k points $\{z_1, \ldots, z_k\}$ shattered by $\mathcal{A}$ [Vapnik and Chervonenkis, 1971]; if no such largest k exists, we define $\mathrm{vc}(\mathcal{A}) = \infty$. For a set $\mathcal{G}$ of real-valued functions, denote by $\mathrm{vc}(\mathcal{G})$ the VC dimension of the collection $\{\{(x, y) : y < g(x)\} : g \in \mathcal{G}\}$ of subgraphs of functions in $\mathcal{G}$ (called the pseudo-dimension [Haussler, 1992, Pollard, 1990]); to simplify the statement of results below, we adopt the convention that when the VC dimension of this collection is 0, we let $\mathrm{vc}(\mathcal{G}) = 1$. A set $\mathcal{G}$ is said to be a VC subgraph class if $\mathrm{vc}(\mathcal{G}) < \infty$ [van der Vaart and Wellner, 1996].

Because we are interested in results concerning values of $\mathrm{R}_\ell(h) - \mathrm{R}_\ell(h^*)$, for functions h in certain subsets $\mathcal{H} \subseteq [\mathcal{F}]$, we will formulate results below in terms of $\mathrm{vc}(\mathcal{G}_\mathcal{H})$, for $\mathcal{G}_\mathcal{H}$ defined as above. Depending on certain properties of ℓ, these results can often be restated directly in terms of $\mathrm{vc}(\mathcal{H})$; for instance, this is true when ℓ is monotone, since $\mathrm{vc}(\mathcal{G}_\mathcal{H}) \leq \mathrm{vc}(\mathcal{H})$ in that case [Dudley, 1987, Haussler, 1992, Nolan and Pollard, 1987].

The following is a well-known result for VC subgraph classes [see e.g., van der Vaart and Wellner, 1996], derived from the works of Pollard [1984] and Haussler [1992].

Lemma 12.14. *For any $\mathcal{G} \subseteq \mathcal{G}^*$, for any measurable $\mathrm{F} \geq \mathrm{F}(\mathcal{G})$, for any distribution Q such that*

$\|\mathrm{F}\|_Q > 0$, *for any* $\varepsilon \in (0,1)$,

$$\mathcal{N}(\varepsilon\|\mathrm{F}\|_Q, \mathcal{G}, L_2(Q)) \leq A(\mathcal{G})\left(\frac{1}{\varepsilon}\right)^{2\mathrm{vc}(\mathcal{G})}.$$

where $A(\mathcal{G}) \lesssim (\mathrm{vc}(\mathcal{G})+1)(16e)^{\mathrm{vc}(\mathcal{G})}$.

In particular, Lemma 12.14 implies that any $\mathcal{G} \subseteq \mathcal{G}^*$ has, $\forall \sigma \in (0,1]$,

$$J(\sigma, \mathcal{G}, \mathrm{F}) \leq \int_0^\sigma \sqrt{\ln(eA(\mathcal{G})) + 2\mathrm{vc}(\mathcal{G})\ln(1/\varepsilon)}\,\mathrm{d}\varepsilon \tag{12.23}$$

$$\leq 2\sigma\sqrt{\ln(eA(\mathcal{G}))} + \sqrt{8\mathrm{vc}(\mathcal{G})}\int_0^\sigma \sqrt{\ln(1/\varepsilon)}\,\mathrm{d}\varepsilon$$

$$= 2\sigma\sqrt{\ln(eA(\mathcal{G}))} + \sigma\sqrt{8\mathrm{vc}(\mathcal{G})\ln(1/\sigma)} + \sqrt{2\pi\mathrm{vc}(\mathcal{G})}\mathrm{erfc}\left(\sqrt{\ln(1/\sigma)}\right).$$

Since $\mathrm{erfc}(x) \leq \exp\{-x^2\}$ for all $x \geq 0$, (12.23) implies $\forall \sigma \in (0,1]$,

$$J(\sigma, \mathcal{G}, \mathrm{F}) \lesssim \sigma\sqrt{\mathrm{vc}(\mathcal{G})\mathrm{Log}(1/\sigma)}. \tag{12.24}$$

Applying these observations to bound $J(\sigma, \mathcal{G}_{\mathcal{H},P}, \mathrm{F})$ for $\mathcal{H} \subseteq [\mathcal{F}]$ and $\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})$, noting $J(\sigma, \mathcal{G}_\mathcal{H}, \mathrm{F}) = J(\sigma, \mathcal{G}_{\mathcal{H},P}, \mathrm{F})$ and $\mathrm{vc}(\mathcal{G}_{\mathcal{H},P}) = \mathrm{vc}(\mathcal{G}_\mathcal{H})$, and plugging the resulting bound into (12.19) yields the following well-known bound on $\mathring{\phi}_\ell^{(1)}$ due to Giné and Koltchinskii [2006]. For any $m \in \mathbb{N}$ and $\sigma > 0$,

$$\mathring{\phi}_\ell^{(1)}(\sigma, \mathcal{H}; m, P)$$

$$\lesssim \inf_{\lambda \geq \sigma} \lambda\sqrt{\frac{\mathrm{vc}(\mathcal{G}_\mathcal{H})\mathrm{Log}\left(\frac{\|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P}{\lambda}\right)}{m}} + \frac{\mathrm{vc}(\mathcal{G}_\mathcal{H})\bar{\ell}\mathrm{Log}\left(\frac{\|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P}{\lambda}\right)}{m}. \tag{12.25}$$

Specifically, to arrive at (12.25), we relaxed the $\inf_{\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})}$ in (12.19) by taking $\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{H},P})$ such that $\|\mathrm{F}\|_P = \max\{\sigma, \|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P\}$, thus maintaining $\lambda/\|\mathrm{F}\|_P \in (0,1]$ for the minimizing λ value, so that (12.24) remains valid; we also made use of the fact that $\mathrm{Log} \geq 1$, which gives us $\mathrm{Log}(\|\mathrm{F}\|_P/\lambda) = \mathrm{Log}(\|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P/\lambda)$ for this case.

In particular, (12.25) implies

$$\ddot{\mathrm{M}}_\ell(\gamma_1, \gamma_2; \mathcal{H}, P)$$

$$\lesssim \inf_{\sigma \geq \mathrm{D}_\ell([\mathcal{H}](\gamma_2;\ell,P);P)} \left(\frac{\sigma^2}{\gamma_1^2} + \frac{\bar{\ell}}{\gamma_1}\right)\mathrm{vc}(\mathcal{G}_\mathcal{H})\mathrm{Log}\left(\frac{\|\mathrm{F}(\mathcal{G}_{\mathcal{H},P})\|_P}{\sigma}\right). \tag{12.26}$$

249

Following Giné and Koltchinskii [2006], for $r > 0$, define $\mathrm{B}_{\mathcal{H},P}(h^*{}_P, r; \ell) = \{g \in \mathcal{H} : \mathrm{D}_\ell(g, h^*{}_P; P)^2 \le r\}$, and for $r_0 \ge 0$, define

$$\tau_\ell(r_0; \mathcal{H}, P) = \sup_{r > r_0} \frac{\left\| \mathrm{F}\left(\mathcal{G}_{\mathrm{B}_{\mathcal{H},P}(h^*{}_P, r; \ell), P}\right) \right\|_P^2}{r} \vee 1.$$

When $P = \mathcal{P}_{XY}$, abbreviate this as $\tau_\ell(r_0; \mathcal{H}) = \tau_\ell(r_0; \mathcal{H}, \mathcal{P}_{XY})$, and when $\mathcal{H} = \mathcal{F}$, further abbreviate $\tau_\ell(r_0) = \tau_\ell(r_0; \mathcal{F}, \mathcal{P}_{XY})$. For $\lambda > 0$, when $h^*{}_P \in \mathcal{H}$ and P satisfies Condition 12.11, (12.26) implies that,

$$\sup_{\gamma \ge \lambda} \ddot{\mathrm{M}}_\ell(\gamma/(4\tilde{K}), \gamma; \mathcal{H}(\gamma; \ell, P), P)$$

$$\lesssim \left(\frac{b}{\lambda^{2-\beta}} + \frac{\bar{\ell}}{\lambda}\right) \mathrm{vc}(\mathcal{G}_\mathcal{H}) \mathrm{Log}\left(\tau_\ell\left(b\lambda^\beta; \mathcal{H}, P\right)\right). \quad (12.27)$$

Combining this observation with (12.6), (12.8), (12.9), (12.10), and Theorem 12.6, we arrive at a result for the sample complexity of empirical ℓ-risk minimization with a general VC subgraph class under Conditions 12.10 and 12.11. Specifically, for $\mathfrak{s} : (0, \infty)^2 \to [1, \infty)$, when $h^* \in \mathcal{F}$, (12.6) implies that

$$\bar{\mathrm{M}}_\ell(\Gamma_\ell(\varepsilon); \mathcal{F}, \mathcal{P}_{XY}, \mathfrak{s}) \le \tilde{\mathrm{M}}_\ell(\Gamma_\ell(\varepsilon); \mathcal{F}, \mathcal{P}_{XY}, \mathfrak{s})$$

$$= \sup_{\gamma \ge \Gamma_\ell(\varepsilon)} \tilde{\mathrm{M}}_\ell(\gamma/2, \gamma; \mathcal{F}(\gamma; \ell), \mathcal{P}_{XY}, \mathfrak{s}(\Gamma_\ell(\varepsilon), \gamma))$$

$$\le \sup_{\gamma \ge \Gamma_\ell(\varepsilon)} \mathring{\mathrm{M}}_\ell(\gamma/2, \gamma; \mathcal{F}(\gamma; \ell), \mathcal{P}_{XY}, \mathfrak{s}(\Gamma_\ell(\varepsilon), \gamma)). \quad (12.28)$$

Supposing $\mathcal{P}_{XY}$ satisfies Conditions 12.10 and 12.11, applying (12.8), (12.9), and (12.27) to (12.28), and taking $\mathfrak{s}(\lambda, \gamma) = \mathrm{Log}\left(\frac{12\gamma}{\lambda\delta}\right)$, we arrive at the following theorem, which is implicit in the work of Giné and Koltchinskii [2006].

Theorem 12.15. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 12.10 and Condition 12.11, ℓ is classification-calibrated, $h^* \in \mathcal{F}$, and Ψ_ℓ is as in (12.15), then for any $\varepsilon \in (0, 1)$, letting $\tau_\ell = \tau_\ell\left(b\Psi_\ell(\varepsilon)^\beta\right)$, for any $m \in \mathbb{N}$ with*

$$m \ge c \left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right) \left(\mathrm{vc}(\mathcal{G}_\mathcal{F}) \mathrm{Log}\left(\tau_\ell\right) + \mathrm{Log}\left(1/\delta\right)\right), \quad (12.29)$$

with probability at least $1 - \delta$, $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ produces $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \le \varepsilon$.

As noted by Giné and Koltchinskii [2006], in the special case when ℓ is itself the 0-1 loss, the bound in Theorem 12.15 simplifies quite nicely, since in that case $\|F(\mathcal{G}_{B_{\mathcal{F},\mathcal{P}_{XY}}(h^*,r;\ell),\mathcal{P}_{XY}})\|^2_{\mathcal{P}_{XY}} = \mathcal{P}(DIS(B(h^*,r)))$, so that $\tau_\ell(r_0) = \theta(r_0)$; in this case, we also have $vc(\mathcal{G}_\mathcal{F}) \leq vc(\mathcal{F})$ and $\Psi_\ell(\varepsilon) = \varepsilon/2$, and we can take $\beta = \alpha$ and $b = a$, so that it suffices to have

$$m \geq ca\varepsilon^{\alpha-2}(vc(\mathcal{F})Log(\theta) + Log(1/\delta)), \qquad (12.30)$$

where $\theta = \theta(a\varepsilon^\alpha)$ and $c \in [1,\infty)$ is a universal constant. It is known that this is sometimes the minimax optimal number of samples sufficient for passive learning [Castro and Nowak, 2008, Hanneke, 2011, Raginsky and Rakhlin, 2011].

Next, we turn to the performance of Algorithm 1 under the conditions of Theorem 12.15. Specifically, suppose $\mathcal{P}_{XY}$ satisfies Conditions 12.10 and 12.11, and for $\gamma_0 \geq 0$, define

$$\chi_\ell(\gamma_0) = \sup_{\gamma > \gamma_0} \frac{\mathcal{P}(DIS(B(h^*, a\mathcal{E}_\ell(\gamma)^\alpha)))}{b\gamma^\beta} \vee 1.$$

Note that $\|F(\mathcal{G}_{\mathcal{F}_j,\mathcal{P}_{XY}})\|^2_{\mathcal{P}_{XY}} \leq \bar{\ell}^2 \mathcal{P}(DIS(\mathcal{F}(\mathcal{E}_\ell(2^{2-j});_{01})))$. Also, note that $vc(\mathcal{G}_{\mathcal{F}_j}) \leq vc(\mathcal{G}_{\mathcal{F}(\mathcal{E}_\ell(2^{2-j});01)}) \leq vc(\mathcal{G}_\mathcal{F})$. Thus, by (12.26), for $j_\ell \leq j \leq \lceil \log_2(1/\Psi_\ell(\varepsilon)) \rceil$,

$$\ddot{M}_\ell(2^{-j-2}\tilde{K}^{-1}, 2^{2-j}; \mathcal{F}_j, \mathcal{P}_{XY}) \lesssim (b2^{j(2-\beta)} + \bar{\ell}2^j) vc(\mathcal{G}_\mathcal{F})Log(\chi_\ell(\Psi_\ell(\varepsilon))\bar{\ell}). \qquad (12.31)$$

With a little additional work to define an appropriate $\hat{s}$ function and derive closed-form bounds on the summation in Theorem 12.7, we arrive at the following theorem regarding the performance of Algorithm 1 for VC subgraph classes. For completeness, the remaining technical details of the proof are included in Appendix 12.6

Theorem 12.16. *For a universal constant $c \in [1,\infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 12.10 and Condition 12.11, ℓ is classification-calibrated, $h^* \in \mathcal{F}$, and Ψ_ℓ is as in (12.15), for any $\varepsilon \in (0,1)$, letting $\theta = \theta(a\varepsilon^\alpha)$, $\chi_\ell = \chi_\ell(\Psi_\ell(\varepsilon))$, $A_1 = vc(\mathcal{G}_\mathcal{F})Log(\chi_\ell\bar{\ell}) + Log(1/\delta)$, $B_1 = \min\left\{\frac{1}{1-2^{(\alpha+\beta-2)}}, Log(\bar{\ell}/\Psi_\ell(\varepsilon))\right\}$, and $C_1 = \min\left\{\frac{1}{1-2^{(\alpha-1)}}, Log(\bar{\ell}/\Psi_\ell(\varepsilon))\right\}$, if*

$$u \geq c\left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)A_1 \qquad (12.32)$$

251

and

$$n \geq c\theta a\varepsilon^{\alpha} \left(\frac{b(A_1 + \mathrm{Log}(B_1))B_1}{\Psi_{\ell}(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}(A_1 + \mathrm{Log}(C_1))C_1}{\Psi_{\ell}(\varepsilon)} \right), \qquad (12.33)$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function, Algorithm 1 uses at most u unlabeled samples and makes at most n label requests, and with probability at least $1 - \delta$, returns a function $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$.

To be clear, in specifying B_1 and C_1, we have adopted the convention that $1/0 = \infty$ and $\min\{\infty, x\} = x$ for any $x \in \mathbb{R}$, so that B_1 and C_1 are well-defined even when $\alpha = \beta = 1$, or $\alpha = 1$, respectively. Note that, when $\alpha + \beta < 2$, $B_1 = O(1)$, so that the asymptotic dependence on ε in (12.33) is $O\left(\theta\varepsilon^{\alpha}\Psi_{\ell}(\varepsilon)^{\beta-2}\mathrm{Log}(\chi_{\ell})\right)$, while in the case of $\alpha = \beta = 1$, it is $O\left(\theta\mathrm{Log}(1/\varepsilon)(\mathrm{Log}(\theta) + \mathrm{Log}(\mathrm{Log}(1/\varepsilon)))\right)$. It is likely that the logarithmic and constant factors can be improved in many cases (particularly the $\mathrm{Log}(\chi_{\ell}\bar{\ell})$, B_1, and C_1 factors).

Comparing the result in Theorem 12.16 to Theorem 12.15, we see that the condition on u in (12.32) is almost identical to the condition on m in (12.29), aside from a change in the logarithmic factor, so that the total number of data points needed is roughly the same. However, the number of *labels* indicated by (12.33) may often be significantly smaller than the condition in (12.29), reducing it by a factor of roughly $\theta a\varepsilon^{\alpha}$. This reduction is particularly strong when θ is bounded by a finite constant. Moreover, this is the same *type* of improvement that is known to occur when ℓ is itself the 0-1 loss [Hanneke, 2011], so that in particular these results agree with the existing analysis in this special case, and are therefore sometimes nearly minimax [Hanneke, 2011, Raginsky and Rakhlin, 2011]. Regarding the slight difference between (12.32) and (12.29) from replacing τ_{ℓ} by $\chi_{\ell}\bar{\ell}$, the effect is somewhat mixed, and which of these is smaller may depend on the particular class $\mathcal{F}$ and loss ℓ; we can generally bound χ_{ℓ} as a function of $\theta(a\varepsilon^{\alpha})$, ψ_{ℓ}, a, α, b, and β. In the special case of ℓ equal the 0-1 loss, both τ_{ℓ} and $\chi_{\ell}\bar{\ell}$ are equal to $\theta(a(\varepsilon/2)^{\alpha})$.

We note that the values $\hat{\mathfrak{s}}(m)$ used in the proof of Theorem 12.16 have a direct dependence on the parameters b, β, a, α, and χ_{ℓ}. Such a dependence may be undesirable for many applications, where information about these values is not available. However, one can easily follow this same

proof, taking $\hat{\mathfrak{s}}(m) = \mathrm{Log}\left(\frac{12\log_2(2m)^2}{\delta}\right)$ instead, which only leads to an increase by a $\log\log$ factor: specifically, replacing the factor of A_1 in (12.32), and the factors $(A_1 + \mathrm{Log}(B_1))$ and $(A_1 + \mathrm{Log}(C_1))$ in (12.33), with a factor of $(A_1 + \mathrm{Log}(\mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))))$. It is not clear whether it is always possible to achieve the slightly tighter result of Theorem 12.16 without having direct access to the values b, β, a, α, and χ_ℓ in the algorithm.

As mentioned above, though convenient in the sense that it offers a completely abstract and unified approach, the choice of $\hat{T}_\ell(V; Q, m)$ given by (12.11) may often make Algorithm 1 computationally inefficient. However, for each of the applications studied here, we can relax this $\hat{T}_\ell$ function to a computationally-accessible value, which will then allow the algorithm to be efficient under convexity conditions on the loss and class of functions. In particular, in the present application to VC Subgraph classes, Theorem 12.16 remains valid if we instead define $\hat{T}_\ell$ as follows. If we let $V^{(m)}$ and Q_m denote the sets V and Q upon reaching Step 5 for any given value of m with $\log_2(m) \in \mathbb{N}$ realized in Algorithm 1, then consider defining $\hat{T}_\ell$ in Step 6 inductively by letting $\hat{\gamma}_{m/2} = \frac{8(|Q_{m/2}|\vee 1)}{m}\left(\hat{T}_\ell(V^{(m/2)}; Q_{m/2}, m/2) \wedge \bar{\ell}\right)$ (or $\hat{\gamma}_{m/2} = \bar{\ell}$ if $m = 2$), and taking (with a slight abuse of notation to allow $\hat{T}_\ell$ to depend on sets $V^{(m')}$ and $Q_{m'}$ with $m' < m$)

$$\hat{T}_\ell(V^{(m)}; Q_m, m) =$$

$$c_0 \frac{m/2}{|Q_m| \vee 1} \left(\sqrt{\hat{\gamma}_{m/2}^\beta \frac{b}{m} \left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}(|Q_m| + \hat{\mathfrak{s}}(m))}{mb\hat{\gamma}_{m/2}^\beta} \right) + \hat{\mathfrak{s}}(m) \right)} \right.$$

$$\left. + \frac{\bar{\ell}}{m}\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}(|Q_m| + \hat{\mathfrak{s}}(m))}{mb\hat{\gamma}_{m/2}^\beta} \right) + \hat{\mathfrak{s}}(m) \right) \right), \quad (12.34)$$

for an appropriate universal constant c_0. This value is essentially derived by upper bounding $\frac{m/2}{|Q|\vee 1}\tilde{U}_\ell(V_{\mathrm{DIS}(V)}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m))$ (which is a bound on (12.11) by Lemma 12.4), based on (12.25) and Condition 12.11 (along with a Chernoff bound to argue $|Q_m| \approx \mathcal{P}(\mathrm{DIS}(V))m/2$; since the sample sizes derived for u and n in Theorem 12.16 are based on these relaxations anyway, they remain sufficient (with slight changes to the constant factors) for these relaxed $\hat{T}_\ell$

values. For brevity, we defer a more detailed proof that these values of $\hat{T}_\ell$ suffice to achieve Theorem 12.16 to Appendix 12.7. Note that we have introduced a dependence on b and β in (12.34). These values would indeed be available for some applications, such as when they are derived from Lemma 12.12 when Condition 12.3 is satisfied; however, in other cases, there may be more-favorable values of b and β than given by Lemma 12.12, dependent on the specific $\mathcal{P}_{XY}$ distribution, and in these cases direct observation of these values might not be available. Thus, there remains an interesting open question of whether there exists a function $\hat{T}_\ell(V; Q, m)$, which is efficiently computable (under convexity assumptions) and yet preserves the validity of Theorem 12.16; this same question applies to each of the results below as well.

In the special case when ℓ satisfies Condition 12.3, we can derive a sometimes-stronger result via Corollary 12.9. Specifically, we can combine (12.26), (12.8), (12.9), and Lemma 12.12, to get that if $h^* \in \mathcal{F}$ and Condition 12.3 is satisfied, then for $j \geq j_\ell$ in Corollary 12.9,

$$
\begin{aligned}
&\mathring{\mathrm{M}}_\ell \left(\frac{2^{-j-7}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{2-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, s \right) \\
&\lesssim \left(b \left(2^j \mathcal{P}(\mathcal{U}_j) \right)^{2-\beta} + 2^j \bar{\ell} \mathcal{P}(\mathcal{U}_j) \right) \left(\mathrm{vc}(\mathcal{G}_\mathcal{F}) \mathrm{Log} \left(\bar{\ell} 2^{j\beta} \mathcal{P}(\mathcal{U}_j)^\beta / b \right) + s \right),
\end{aligned}
\tag{12.35}
$$

where b and β are as in Lemma 12.12. Plugging this into Corollary 12.9, with $\hat{\mathfrak{s}}$ defined analogous to that used in the proof of Theorem 12.16, and bounding the summation in the condition for n in Corollary 12.9, we arrive at the following theorem. The details of the proof proceed along similar lines as the proof of Theorem 12.16, and a sketch of the remaining technical details is included in Appendix 12.6.

Theorem 12.17. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 12.10, ℓ is classification-calibrated and satisfies Condition 12.3, $h^* \in \mathcal{F}$, Ψ_ℓ is as in (12.15), and b and β are as in Lemma 12.12, then for any $\varepsilon \in (0, 1)$, letting $\theta = \theta(a\varepsilon^\alpha)$, $A_2 = \mathrm{vc}(\mathcal{G}_\mathcal{F}) \mathrm{Log} \left((\bar{\ell}/b) (a\theta\varepsilon^\alpha / \Psi_\ell(\varepsilon))^\beta \right) + \mathrm{Log}(1/\delta)$, $B_2 = \min \left\{ \frac{1}{1 - 2^{(\alpha-1)(2-\beta)}}, \mathrm{Log} \left(\bar{\ell}/\Psi_\ell(\varepsilon) \right) \right\}$, and $C_2 = \min \left\{ \frac{1}{1 - 2^{(\alpha-1)}}, \mathrm{Log} \left(\bar{\ell}/\Psi_\ell(\varepsilon) \right) \right\}$, if*

$$
u \geq c \left(\frac{b (a\theta\varepsilon^\alpha)^{1-\beta}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)} \right) A_2,
\tag{12.36}
$$

and

$$n \geq c \left(b(A_2 + \text{Log}(B_2)) B_2 \left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^{2-\beta} + \bar{\ell}(A_2 + \text{Log}(C_2)) C_2 \left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right) \right), \qquad (12.37)$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function, Algorithm 1 uses at most u unlabeled samples and makes at most n label requests, and with probability at least $1 - \delta$, returns a function $\hat{h}$ with $\text{er}(\hat{h}) - \text{er}(h^*) \leq \varepsilon$.

Examining the asymptotic dependence on ε in the above result, the sufficient number of unlabeled samples is $O\left(\frac{(\theta\varepsilon^\alpha)^{1-\beta}}{\Psi_\ell(\varepsilon)^{2-\beta}} \text{Log}\left(\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^\beta \right) \right)$, and the sufficient number of label requests is $O\left(\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^{2-\beta} \text{Log}\left(\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^\beta \right) \right)$ in the case that $\alpha < 1$, or $O\left(\theta^{2-\beta} \text{Log}(1/\varepsilon) \text{Log}\left(\theta^\beta \text{Log}(1/\varepsilon) \right) \right)$ in the case that $\alpha = 1$. This is noteworthy in the case $\alpha > 0$ and $r_\ell > 2$, for at least two reasons. First, the number of label requests indicated by this result can often be smaller than that indicated by Theorem 12.16, by a factor of roughly $\tilde{O}\left((\theta\varepsilon^\alpha)^{1-\beta} \right)$; this is particularly interesting when θ is bounded by a finite constant. The second interesting feature of this result is that even the sufficient number of *unlabeled* samples, as indicated by (12.36), can often be smaller than the number of *labeled* samples sufficient for ERM_ℓ, as indicated by Theorem 12.15, again by a factor of roughly $\tilde{O}\left((\theta\varepsilon^\alpha)^{1-\beta} \right)$. This indicates that, in the case of a surrogate loss ℓ satisfying Condition 12.3 with $r_\ell > 2$, when Theorem 12.15 is tight, even if we have complete access to a fully labeled data set, we may still prefer to use Algorithm 1 rather than ERM_ℓ; this is somewhat surprising, since (as (12.37) indicates) we expect Algorithm 1 to ignore the vast majority of the labels in this case. That said, it is not clear whether there exist natural classification-calibrated losses ℓ satisfying Condition 12.3 with $r_\ell > 2$ for which the indicated sufficient size of m in Theorem 12.15 is ever competitive with the known results for methods that directly optimize the empirical 0-1 risk (i.e., Theorem 12.15 with ℓ the 0-1 loss); thus, the improvements in u and n reflected by Theorem 12.17 may simply indicate that Algorithm 1 is, to some extent, *compensating* for a choice of loss ℓ that would otherwise lead to suboptimal label complexities.

We note that, as in Theorem 12.16, the values $\hat{\mathfrak{s}}$ used to obtain this result have a direct dependence on certain values, which are typically not directly accessible in practice: in this

case, a, α, and θ. However, as was the case for Theorem 12.16, we can obtain only slightly worse results by instead taking $\hat{\mathfrak{s}}(m) = \mathrm{Log}\left(\frac{12\log_2(2m)^2}{\delta}\right)$, which again only leads to an increase by a $\log\log$ factor: replacing the factor of A_2 in (12.36), and the factors $(A_2 + \mathrm{Log}(B_2))$ and $(A_2 + \mathrm{Log}(C_2))$ in (12.37), with a factor of $(A_2 + \mathrm{Log}(\mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))))$. As before, it is not clear whether the slightly tighter result of Theorem 12.17 is always available, without requiring direct dependence on these quantities.

As was also true of Theorem 12.16, while the above choice of $\hat{T}_\ell(V; Q, m)$ given by (12.11) provides an elegant unifying perspective, it may often be infeasible to calculate efficiently. However, as was possible in that case, we can define an alternative that is specialized to the conditions of Theorem 12.17, for which the theorem statement remains valid. Specifically, consider instead defining $\hat{T}_\ell$ in Step 6 as

$$
\hat{T}_\ell(V^{(m)}; Q_m, m)
$$
$$
= c_0 \left(\frac{b}{|Q_m| \vee 1} \left(\mathrm{vc}(\mathcal{G}_\mathcal{F}) \mathrm{Log}\left(\frac{\bar{\ell}}{b} \left(\frac{|Q_m|}{b\mathrm{vc}(\mathcal{G}_\mathcal{F})} \right)^{\frac{\beta}{2-\beta}} \right) + \hat{\mathfrak{s}}(m) \right) \right)^{\frac{1}{2-\beta}} \wedge \bar{\ell}, \quad (12.38)
$$

for b and β as in Lemma 12.12, and for an appropriate universal constant c_0. This value is essentially derived by bounding $\tilde{U}_\ell(V; \mathcal{P}_{\mathrm{DIS}(V)}, \hat{\mathfrak{s}}(m))$, which is informative in Step 6 via Lemma 12.4. Since Theorem 12.17 is proven by considering concentration under the conditional distributions $\mathcal{P}_{\mathcal{U}_j}$ via Corollary 12.9, and (12.38) represents the concentration bound one gets from directly applying Lemma 12.4 to the samples from the conditional distribution $\mathcal{P}_{\mathrm{DIS}(V^{(m)})}$, one can show that the conclusions of Theorem 12.17 remain valid for this specification of $\hat{T}_\ell$ in place of (12.11). For brevity, the details of the proof are omitted. Note that, unlike the analogous result for Theorem 12.16 based on (12.34) above, in this case all of the quantities in $\hat{T}_\ell(V; Q, m)$ are directly observable (in particular, b and β), aside from any possible dependence arising in the specification of $\hat{\mathfrak{s}}$.

12.5.5 Entropy Conditions

Next we turn to problems satisfying certain entropy conditions. In particular, the following represent two commonly-studied conditions, which allow for concise statement of results below.

Condition 12.18. *For some $q \geq 1$, $\rho \in (0,1)$, and $\mathrm{F} \geq \mathrm{F}(\mathcal{G}_{\mathcal{F}, \mathcal{P}_{XY}})$, either $\forall \varepsilon > 0$,*

$$\ln \mathcal{N}_{[]}(\varepsilon \|\mathrm{F}\|_{\mathcal{P}_{XY}}, \mathcal{G}_{\mathcal{F}}, L_2(\mathcal{P}_{XY})) \leq q\varepsilon^{-2\rho}, \tag{12.39}$$

or for all finitely discrete P, $\forall \varepsilon > 0$,

$$\ln \mathcal{N}(\varepsilon \|\mathrm{F}\|_P, \mathcal{G}_{\mathcal{F}}, L_2(P)) \leq q\varepsilon^{-2\rho}. \tag{12.40}$$

In particular, note that when $\mathcal{F}$ satisfies Condition 12.18, for $0 \leq \sigma \leq 2\|\mathrm{F}\|_{\mathcal{P}_{XY}}$,

$$\mathring{\phi}_\ell(\sigma, \mathcal{F}; \mathcal{P}_{XY}, m) \lesssim \max\left\{ \frac{\sqrt{q}\|\mathrm{F}\|_{\mathcal{P}_{XY}}^\rho \sigma^{1-\rho}}{(1-\rho)m^{1/2}}, \frac{\bar{\ell}^{\frac{1-\rho}{1+\rho}} q^{\frac{1}{1+\rho}} \|\mathrm{F}\|_{\mathcal{P}_{XY}}^{\frac{2\rho}{1+\rho}}}{(1-\rho)^{\frac{2}{1+\rho}} m^{\frac{1}{1+\rho}}} \right\}. \tag{12.41}$$

Since $\mathrm{D}_\ell([\mathcal{F}]) \leq 2\|\mathrm{F}\|_{\mathcal{P}_{XY}}$, this implies that for any numerical constant $c \in (0,1]$, for every $\gamma \in (0, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 12.11, then

$$\ddot{\mathrm{M}}_\ell(c\gamma, \gamma; \mathcal{F}, \mathcal{P}_{XY}) \lesssim \frac{q\|\mathrm{F}\|_{\mathcal{P}_{XY}}^{2\rho}}{(1-\rho)^2} \max\left\{ b^{1-\rho}\gamma^{\beta(1-\rho)-2}, \bar{\ell}^{1-\rho}\gamma^{-(1+\rho)} \right\}. \tag{12.42}$$

Combined with (12.8), (12.9), (12.10), and Theorem 12.6, taking $\mathfrak{s}(\lambda, \gamma) = \mathrm{Log}\left(\frac{12\gamma}{\lambda\delta}\right)$, we arrive at the following classic result [e.g., Bartlett, Jordan, and McAuliffe, 2006, van der Vaart and Wellner, 1996].

Theorem 12.19. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 12.10 and Condition 12.11, $\mathcal{F}$ and $\mathcal{P}_{XY}$ satisfy Condition 12.18, ℓ is classification-calibrated, $h^* \in \mathcal{F}$, and Ψ_ℓ is as in (12.15), then for any $\varepsilon \in (0,1)$ and m with*

$$m \geq c \frac{q\|\mathrm{F}\|_{\mathcal{P}_{XY}}^{2\rho}}{(1-\rho)^2} \left(\frac{b^{1-\rho}}{\Psi_\ell(\varepsilon)^{2-\beta(1-\rho)}} + \frac{\bar{\ell}^{1-\rho}}{\Psi_\ell(\varepsilon)^{1+\rho}} \right)$$

$$+ c \left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)} \right) \mathrm{Log}\left(\frac{1}{\delta}\right),$$

with probability at least $1 - \delta$, $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ produces $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$.

Next, turning to the analysis of Algorithm 1 under these same conditions, combining (12.42) with (12.8), (12.9), and Theorem 12.7, we have the following result. The details of the proof follow analogously to the proof of Theorem 12.16, and are therefore omitted for brevity.

Theorem 12.20. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 12.10 and Condition 12.11, $\mathcal{F}$ and $\mathcal{P}_{XY}$ satisfy Condition 12.18, ℓ is classification-calibrated, $h^* \in \mathcal{F}$, and Ψ_ℓ is as in (12.15), then for any $\varepsilon \in (0, 1)$, letting B_1 and C_1 be as in Theorem 12.16, $B_3 = \min\left\{\frac{1}{1-2^{(\alpha+\beta(1-\rho)-2)}}, \mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))\right\}$, $C_3 = \min\left\{\frac{1}{1-2^{(\alpha-(1+\rho))}}, \mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon))\right\}$, and $\theta = \theta(a\varepsilon^\alpha)$, if*

$$
u \geq c\frac{q\|\mathrm{F}\|_{\mathcal{P}_{XY}}^{2\rho}}{(1-\rho)^2}\left(\frac{b^{1-\rho}}{\Psi_\ell(\varepsilon)^{2-\beta(1-\rho)}} + \frac{\bar{\ell}^{1-\rho}}{\Psi_\ell(\varepsilon)^{1+\rho}}\right)
$$

$$
+ c\left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\mathrm{Log}\left(\frac{1}{\delta}\right) \quad (12.43)
$$

and

$$
n \geq c\theta a\varepsilon^\alpha\frac{q\|\mathrm{F}\|_{\mathcal{P}_{XY}}^{2\rho}}{(1-\rho)^2}\left(\frac{b^{1-\rho}B_3}{\Psi_\ell(\varepsilon)^{2-\beta(1-\rho)}} + \frac{\bar{\ell}^{1-\rho}C_3}{\Psi_\ell(\varepsilon)^{1+\rho}}\right)
$$

$$
+ c\theta a\varepsilon^\alpha\left(\frac{bB_1\mathrm{Log}(B_1/\delta)}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}C_1\mathrm{Log}(C_1/\delta)}{\Psi_\ell(\varepsilon)}\right), \quad (12.44)
$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function, Algorithm 1 uses at most u unlabeled samples and makes at most n label requests, and with probability at least $1 - \delta$, returns a function $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$.

The sufficient size of u in Theorem 12.20 is essentially identical (up to the constant factors) to the number of labels sufficient for ERM_ℓ to achieve the same, as indicated by Theorem 12.19. In particular, the dependence on ε in these results is $O\left(\Psi_\ell(\varepsilon)^{\beta(1-\rho)-2}\right)$. On the other hand, when $\theta(\varepsilon^\alpha) = o(\varepsilon^{-\alpha})$, the sufficient size of n in Theorem 12.20 *does* reflect an improvement in the number of labels indicated by Theorem 12.19, by a factor with dependence on ε of $O\left(\theta\varepsilon^\alpha\right)$.

As before, in the special case when ℓ satisfies Condition 12.3, we can derive sometimes stronger results via Corollary 12.9. In this case, we will distinguish between the cases of (12.40) and (12.39), as we find a slightly stronger result for the former.

First, suppose (12.40) is satisfied for all finitely discrete P and all $\varepsilon > 0$, with $\mathrm{F} \leq \bar{\ell}$. Then following the derivation of (12.42) above, combined with (12.9), (12.8), and Lemma 12.12, for values of $j \geq j_\ell$ in Corollary 12.9,

$$
\mathring{\mathrm{M}}_\ell \left(\frac{2^{-j-7}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{2-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, s \right)
$$
$$
\lesssim \frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2} \left(b^{1-\rho} \left(2^j \mathcal{P}(\mathcal{U}_j) \right)^{2-\beta(1-\rho)} + \bar{\ell}^{1-\rho} \left(2^j \mathcal{P}(\mathcal{U}_j) \right)^{1+\rho} \right)
$$
$$
+ \left(b \left(2^j \mathcal{P}(\mathcal{U}_j) \right)^{2-\beta} + \bar{\ell} 2^j \mathcal{P}(\mathcal{U}_j) \right) s,
$$

where q and ρ are from Lemma 12.12. This immediately leads to the following result by reasoning analogous to the proof of Theorem 12.17.

Theorem 12.21. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 12.10, ℓ is classification-calibrated and satisfies Condition 12.3, $h^* \in \mathcal{F}$, Ψ_ℓ is as in (12.15), b and β are as in Lemma 12.12, and (12.40) is satisfied for all finitely discrete P and all $\varepsilon > 0$, with $\mathrm{F} \leq \bar{\ell}$, then for any $\varepsilon \in (0, 1)$, letting B_2 and C_2 be as in Theorem 12.17, $B_4 = \min \left\{ \frac{1}{1 - 2^{(\alpha-1)(2-\beta(1-\rho))}}, \mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon)) \right\}$, $C_4 = \min \left\{ \frac{1}{1 - 2^{(\alpha-1)(1+\rho)}}, \mathrm{Log}(\bar{\ell}/\Psi_\ell(\varepsilon)) \right\}$, and $\theta = \theta\left(a\varepsilon^\alpha\right)$, if*

$$
u \geq c \left(\frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2} \right) \left(\left(\frac{b^{1-\rho}}{\Psi_\ell(\varepsilon)} \right) \left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^{1-\beta(1-\rho)} + \left(\frac{\bar{\ell}^{1-\rho}}{\Psi_\ell(\varepsilon)} \right) \left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^\rho \right)
$$
$$
+ c \left(\left(\frac{b}{\Psi_\ell(\varepsilon)} \right) \left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^{1-\beta} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)} \right) \mathrm{Log}(1/\delta)
$$

and

$$
n \geq c \left(\frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2} \right) \left(B_4 b^{1-\rho} \left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^{2-\beta(1-\rho)} + C_4 \bar{\ell}^{1-\rho} \left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^{1+\rho} \right)
$$
$$
+ c \left(B_2 \mathrm{Log}(B_2/\delta) b \left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right)^{2-\beta} + C_2 \mathrm{Log}(C_2/\delta) \bar{\ell} \left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)} \right) \right),
$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function, Algorithm 1 uses at most u unlabeled samples and makes at most n label requests, and with probability at least $1 - \delta$, returns a function $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$.

Compared to Theorem 12.20, in terms of the asymptotic dependence on ε, the sufficient sizes for both u and n here may be smaller by a factor of $O\left((\theta\varepsilon^\alpha)^{1-\beta(1-\rho)}\right)$, which sometimes represents a significant refinement, particularly when θ is much smaller than $\varepsilon^{-\alpha}$. In particular, as was the case in Theorem 12.17, when $\theta(\varepsilon) = o(1/\varepsilon)$, the size of u indicated by Theorem 12.21 is smaller than the known results for $\mathrm{ERM}_\ell(\mathcal{F}, \mathscr{Z}_m)$ from Theorem 12.19.

The case where (12.39) is satisfied can be treated similarly, though the result we obtain here is slightly weaker. Specifically, for simplicity suppose (12.39) is satisfied with $\mathrm{F} = \bar{\ell}$ constant. In this case, we have $\bar{\ell} \geq \mathrm{F}(\mathcal{G}_{\mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}})$ as well, while $\mathcal{N}_{[]}(\varepsilon\bar{\ell}, \mathcal{G}_{\mathcal{F}_j}, L_2(\mathcal{P}_{\mathcal{U}_j})) = \mathcal{N}_{[]}(\varepsilon\bar{\ell}\sqrt{\mathcal{P}(\mathcal{U}_j)}, \mathcal{G}_{\mathcal{F}_j}, L_2(\mathcal{P}_{XY}))$, which is no larger than $\mathcal{N}_{[]}(\varepsilon\bar{\ell}\sqrt{\mathcal{P}(\mathcal{U}_j)}, \mathcal{G}_{\mathcal{F}}, L_2(\mathcal{P}_{XY}))$, so that $\mathcal{F}_j$ and $\mathcal{P}_{\mathcal{U}_j}$ also satisfy (12.39) with $\mathrm{F} = \bar{\ell}$; specifically,

$$\ln \mathcal{N}_{[]}\left(\varepsilon\bar{\ell}, \mathcal{G}_{\mathcal{F}_j}, L_2(\mathcal{P}_{\mathcal{U}_j})\right) \leq q\mathcal{P}(\mathcal{U}_j)^{-\rho}\varepsilon^{-2\rho}.$$

Thus, based on (12.42), (12.8), (12.9), and Lemma 12.12, we have that if $h^* \in \mathcal{F}$ and Condition 12.3 is satisfied, then for $j \geq j_\ell$ in Corollary 12.9,

$$
\mathring{\mathrm{M}}_\ell\left(\frac{2^{-j-7}}{\mathcal{P}(\mathcal{U}_j)}, \frac{2^{2-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, s\right)
$$
$$
\lesssim \left(\frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2}\right) \mathcal{P}(\mathcal{U}_j)^{-\rho}\left(b^{1-\rho}\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta(1-\rho)} + \bar{\ell}^{1-\rho}\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{1+\rho}\right)
$$
$$
+ \left(b\left(2^j\mathcal{P}(\mathcal{U}_j)\right)^{2-\beta} + \bar{\ell}2^j\mathcal{P}(\mathcal{U}_j)\right) s,
$$

where b and β are as in Lemma 12.12. Combining this with Corollary 12.9 and reasoning analogously to the proof of Theorem 12.17, we have the following result.

Theorem 12.22. *For a universal constant $c \in [1, \infty)$, if $\mathcal{P}_{XY}$ satisfies Condition 12.10, ℓ is classification-calibrated and satisfies Condition 12.3, $h^* \in \mathcal{F}$, Ψ_ℓ is as in (12.15), b and β are as in Lemma 12.12, and (12.39) is satisfied with $\mathrm{F} = \bar{\ell}$ constant, then for any $\varepsilon \in (0, 1)$, letting B_2 and C_2 be as in Theorem 12.17, $B_5 = \min\left\{\frac{1}{1-2^{(\alpha-1)(2-\beta(1-\rho))-\alpha\rho}}, \mathrm{Log}\left(\frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\right\}$, $C_5 =$*

$\min\left\{\frac{1}{1-2^{\alpha-1-\rho}}, \mathrm{Log}\left(\frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\right\}$, *and* $\theta = \theta\left(a\varepsilon^\alpha\right)$, *if*

$$u \geq c\left(\frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2}\right)\left(\left(\frac{b^{1-\rho}}{\Psi_\ell(\varepsilon)^{1+\rho}}\right)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{(1-\beta)(1-\rho)} + \frac{\bar{\ell}^{1-\rho}}{\Psi_\ell(\varepsilon)^{1+\rho}}\right)$$

$$+ c\left(\left(\frac{b}{\Psi_\ell(\varepsilon)}\right)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1-\beta} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\mathrm{Log}(1/\delta)$$

and

$$n \geq c\left(\frac{q\bar{\ell}^{2\rho}}{(1-\rho)^2}\right)\left(\left(\frac{B_5 b^{1-\rho}}{\Psi_\ell(\varepsilon)^\rho}\right)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1+(1-\beta)(1-\rho)} + \frac{C_5\bar{\ell}^{1-\rho}a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)^{1+\rho}}\right)$$

$$+ c\left(bB_2\mathrm{Log}(B_2/\delta)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta} + \bar{\ell}C_2\mathrm{Log}(C_2/\delta)\left(\frac{a\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)\right),$$

then, with arguments ℓ, u, and n, and an appropriate $\hat{\mathfrak{s}}$ function, Algorithm 1 uses at most u unlabeled samples and makes at most n label requests, and with probability at least $1 - \delta$, returns a function $\hat{h}$ with $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$.

In this case, compared to Theorem 12.20, in terms of the asymptotic dependence on ε, the sufficient sizes for both u and n here may be smaller by a factor of $O\left((\theta\varepsilon^\alpha)^{(1-\beta)(1-\rho)}\right)$, which may sometimes be significant, though not quite as dramatic a refinement as we found under (12.40) in Theorem 12.21. As with Theorem 12.21, when $\theta(\varepsilon) = o(1/\varepsilon)$, the size of u indicated by Theorem 12.22 is smaller than the known results for $\mathrm{ERM}_\ell(\mathcal{F}, \mathcal{Z}_m)$ from Theorem 12.19.

12.5.6 Remarks on VC Major and VC Hull Classes

Another widely-studied family of function classes includes *VC Major* classes. Specifically, we say $\mathcal{G}$ is a VC Major class with index d if $d = \mathrm{vc}(\{\{z : g(z) \geq t\} : g \in \mathcal{G}, t \in \mathbb{R}\}) < \infty$. We can derive results for VC Major classes, analogously to the above, as follows. For brevity, we leave many of the details as an exercise for the reader. For any VC Major class $\mathcal{G} \subseteq \mathcal{G}^*$ with index d, by reasoning similar to that of Giné and Koltchinskii [2006], one can show that if $\mathrm{F} = \bar{\ell}\mathbb{I}_\mathcal{U} \geq \mathrm{F}(\mathcal{G})$ for some measurable $\mathcal{U} \subseteq \mathcal{X} \times \mathcal{Y}$, then for any distribution P and $\varepsilon > 0$,

$$\ln\mathcal{N}\left(\varepsilon\|\mathrm{F}\|_P, \mathcal{G}, L_2(P)\right) \lesssim \frac{d}{\varepsilon}\log\left(\frac{\bar{\ell}}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon}\right).$$

This implies that for $\mathcal{F}$ a VC Major class, and ℓ classification-calibrated and either nonincreasing or Lipschitz, if $h^* \in \mathcal{F}$ and $\mathcal{P}_{XY}$ satisfies Condition 12.10 and Condition 12.11, then the conditions of Theorem 12.7 can be satisfied with the probability bound being at least $1 - \delta$, for some $u = \tilde{O}\left(\frac{\theta^{1/2}\varepsilon^{\alpha/2}}{\Psi_\ell(\varepsilon)^{2-\beta/2}} + \Psi_\ell(\varepsilon)^{\beta-2}\right)$ and $n = \tilde{O}\left(\frac{\theta^{3/2}\varepsilon^{3\alpha/2}}{\Psi_\ell(\varepsilon)^{2-\beta/2}} + \theta\varepsilon^\alpha\Psi_\ell(\varepsilon)^{\beta-2}\right)$, where $\theta = \theta(a\varepsilon^\alpha)$, and $\tilde{O}(\cdot)$ hides logarithmic and constant factors. Under Condition 12.3, with β as in Lemma 12.12, the conditions of Corollary 12.9 can be satisfied with the probability bound being at least $1 - \delta$, for some $u = \tilde{O}\left(\left(\frac{1}{\Psi_\ell(\varepsilon)}\right)\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1-\beta/2}\right)$ and $n = \tilde{O}\left(\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta/2}\right)$.

For example, for $\mathcal{X} = [0, 1]$ and $\mathcal{F}$ the class of all nondecreasing functions mapping $\mathcal{X}$ to $[-1, 1]$, $\mathcal{F}$ is a VC Major class with index 1, and $\theta(0) \leq 2$ for all distributions $\mathcal{P}$. Thus, for instance, if η is nondecreasing and ℓ is the quadratic loss, then $h^* \in \mathcal{F}$, and Algorithm 1 achieves excess error rate ε with high probability for some $u = \tilde{O}\left(\varepsilon^{2\alpha-3}\right)$ and $n = \tilde{O}\left(\varepsilon^{3(\alpha-1)}\right)$.

VC Major classes are contained in special types of *VC Hull* classes, which are more generally defined as follows. Let $\mathbb{C}$ be a VC Subgraph class of functions on $\mathcal{X}$, with bounded envelope, and for $B \in (0, \infty)$, let $\mathcal{F} = B\text{conv}(\mathbb{C}) = \left\{x \mapsto B\sum_j \lambda_j h_j(x) : \sum_j |\lambda_j| \leq 1, h_j \in \mathbb{C}\right\}$ denote the scaled symmetric convex hull of $\mathbb{C}$; then $\mathcal{F}$ is called a VC Hull class. For instance, these spaces are often used in conjunction with the popular AdaBoost learning algorithm. One can derive results for VC Hull classes following analogously to the above, using established bounds on the uniform covering numbers of VC Hull classes [see van der Vaart and Wellner, 1996, Corollary 2.6.12], and noting that for any VC Hull class $\mathcal{F}$ with envelope function F, and any $\mathcal{U} \subseteq \mathcal{X}$, $\mathcal{F}_\mathcal{U}$ is also a VC Hull class, with envelope function $F\mathbb{I}_\mathcal{U}$. Specifically, one can use these observations to derive the following results. For a VC Hull class $\mathcal{F} = B\text{conv}(\mathbb{C})$ with $d = 2\text{vc}(\mathbb{C})$, if ℓ is classification-calibrated and Lipschitz, $h^* \in \mathcal{F}$, and $\mathcal{P}_{XY}$ satisfies Condition 12.10 and Condition 12.11, then the conditions of Theorem 12.7 can be satisfied with the probability bound being at least $1 - \delta$, for some $u = \tilde{O}\left((\theta\varepsilon^\alpha)^{\frac{d}{d+2}}\Psi_\ell(\varepsilon)^{\frac{2\beta}{d+2}-2}\right)$ and $n = \tilde{O}\left((\theta\varepsilon^\alpha)^{\frac{2d+2}{d+2}}\Psi_\ell(\varepsilon)^{\frac{2\beta}{d+2}-2}\right)$. Under Condition 12.3, with β as in Lemma 12.12, the conditions of Corollary 12.9 can be satisfied with the probability bound being at least $1 - \delta$, for some $u = \tilde{O}\left(\left(\frac{1}{\Psi_\ell(\varepsilon)}\right)\left(\frac{\theta\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{1-\frac{2\beta}{d+2}}\right)$ and

$n = \tilde{O}\left(\left(\frac{\theta \varepsilon^{\alpha}}{\Psi_{\ell}(\varepsilon)}\right)^{2-\frac{2\beta}{d+2}}\right)$. However, it is not clear whether these results for VC Hull classes have any practical implications, since we do not know of any examples of VC Hull classes where these results reflect an improvement over a more direct analysis of ERM_{ℓ} for these scenarios.

12.6 Proofs

Proof of Theorem 12.7. Fix any $\varepsilon \in (0, 1)$, $s \in [1, \infty)$, values u_j satisfying (12.12), and consider running Algorithm 1 with values of u and n satisfying the conditions specified in Theorem 12.7. The proof has two main components: first, showing that, with high probability, $h^* \in V$ is maintained as an invariant, and second, showing that, with high probability, the set V will be sufficiently reduced to provide the guarantee on $\hat{h}$ after at most the stated number of label requests, given the value of u is as large as stated. Both of these components are served by the following application of Lemma 12.4.

Let S denote the set of values of m obtained in Algorithm 1 for which $\log_2(m) \in \mathbb{N}$. For each $m \in S$, let $V^{(m)}$ and Q_m denote the values of V and Q (respectively) upon reaching Step 5 on the round that Algorithm 1 obtains that value of m, and let $\tilde{V}^{(m)}$ denote the value of V upon completing Step 6 on that round; also denote $D_m = \mathrm{DIS}(V^{(m)})$ and $\mathcal{L}_m = \{(1 + m/2, Y_{1+m/2}), \ldots, (m, Y_m)\}$, and define $\tilde{V}^{(1)} = \mathcal{F}$ and $D_1 = \mathrm{DIS}(\mathcal{F})$.

Consider any $m \in S$, and note that $\forall h, g \in V^{(m)}$,

$$(|Q_m| \vee 1)\left(\mathrm{R}_{\ell}(h; Q_m) - \mathrm{R}_{\ell}(g; Q_m)\right)$$

$$= \frac{m}{2}\left(\mathrm{R}_{\ell}(h_{D_m}; \mathcal{L}_m) - \mathrm{R}_{\ell}(g_{D_m}; \mathcal{L}_m)\right), \quad (12.45)$$

and furthermore that

$$(|Q_m| \vee 1)\hat{U}_{\ell}(V^{(m)}; Q_m, \hat{\mathfrak{s}}(m)) = \frac{m}{2}\hat{U}_{\ell}(V_{D_m}^{(k)}; \mathcal{L}_m, \hat{\mathfrak{s}}(m)). \quad (12.46)$$

Applying Lemma 12.4 under the conditional distribution given $V^{(m)}$, combined with the law of total probability, we have that, for every $m \in \mathbb{N}$ with $\log_2(m) \in \mathbb{N}$, on an event of probability

at least $1 - 6e^{-\hat{\mathfrak{s}}(m)}$, if $h^* \in V^{(m)}$ and $m \in S$, then letting $\hat{U}_m = \hat{U}_\ell \left(V_{D_m}^{(m)}; \mathcal{L}_m, \hat{\mathfrak{s}}(m) \right)$, every $h_{D_m} \in V_{D_m}^{(m)}$ has

$$\mathrm{R}_\ell(h_{D_m}) - \mathrm{R}_\ell(h^*) < \mathrm{R}_\ell(h_{D_m}; \mathcal{L}_m) - \mathrm{R}_\ell(h^*; \mathcal{L}_m) + \hat{U}_m, \tag{12.47}$$

$$\mathrm{R}_\ell(h_{D_m}; \mathcal{L}_m) - \min_{g_{D_m} \in V_{D_m}^{(m)}} \mathrm{R}_\ell(g_{D_m}; \mathcal{L}_m) < \mathrm{R}_\ell(h_{D_m}) - \mathrm{R}_\ell(h^*) + \hat{U}_m, \tag{12.48}$$

and furthermore

$$\hat{U}_m < \tilde{U}_\ell \left(V_{D_m}^{(m)}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m) \right). \tag{12.49}$$

By a union bound, on an event of probability at least $1 - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)}$, for every $m \in S$ with $m \le u_{j_\varepsilon}$ and $h^* \in V^{(m)}$, the inequalities (12.47), (12.48), and (12.49) hold. Call this event E.

In particular, note that on the event E, for any $m \in S$ with $m \le u_{j_\varepsilon}$ and $h^* \in V^{(m)}$, since $h^*{}_{D_m} = h^*$, (12.45), (12.48), and (12.46) imply

$$
(|Q_m| \vee 1) \left(\mathrm{R}_\ell(h^*; Q_m) - \inf_{g \in V^{(m)}} \mathrm{R}_\ell(g; Q_m) \right)
$$
$$
= \frac{m}{2} \left(\mathrm{R}_\ell(h^*; \mathcal{L}_m) - \inf_{g_{D_m} \in V_{D_m}^{(m)}} \mathrm{R}_\ell(g_{D_m}; Q_m) \right)
$$
$$
< \frac{m}{2} \hat{U}_m = (|Q_m| \vee 1)\hat{U}_\ell(V^{(m)}; Q_m, \hat{\mathfrak{s}}(m)),
$$

so that $h^* \in \tilde{V}^{(m)}$ as well. Since $h^* \in V^{(2)}$, and every $m \in S$ with $m > 2$ has $V^{(m)} = \tilde{V}^{(m/2)}$, by induction we have that, on the event E, every $m \in S$ with $m \le u_{j_\varepsilon}$ has $h^* \in V^{(m)}$ and $h^* \in \tilde{V}^{(m)}$; this also implies that (12.47), (12.48), and (12.49) all hold for these values of m on the event E.

We next prove by induction that, on the event E, $\forall j \in \{j_\ell - 2, j_\ell - 1, j_\ell, \ldots, j_\varepsilon\}$, if $u_j \in S \cup \{1\}$, then $\tilde{V}_{D_{u_j}}^{(u_j)} \subseteq [\mathcal{F}](2^{-j}; \ell)$ and $\tilde{V}^{(u_j)} \subseteq \mathcal{F}(\mathcal{E}_\ell(2^{-j}); \text{o1})$. This claim is trivially satisfied for $j \in \{j_\ell - 2, j_\ell - 1\}$, since in that case $[\mathcal{F}](2^{-j}; \ell) = [\mathcal{F}] \supseteq \tilde{V}_{D_{u_j}}^{(u_j)}$ and $\mathcal{F}(\mathcal{E}_\ell(2^{-j}); \text{o1}) = \mathcal{F}$, so that these values can serve as our base case. Now take as an inductive hypothesis that, for some $j \in \{j_\ell, \ldots, j_\varepsilon\}$, if $u_{j-2} \in S \cup \{1\}$, then on the event E, $\tilde{V}_{D_{u_{j-2}}}^{(u_{j-2})} \subseteq [\mathcal{F}](2^{2-j}; \ell)$ and

$\tilde{V}^{(u_{j-2})} \subseteq \mathcal{F}\left(\mathcal{E}_\ell(2^{2-j}); {}_{01}\right)$, and suppose the event E occurs. If $u_j \notin S$, the claim is trivially satisfied; otherwise, suppose $u_j \in S$, which further implies $u_{j-2} \in S \cup \{1\}$. Since $u_j \leq u_{j_\varepsilon}$, for any $h \in \tilde{V}^{(u_j)}$, (12.47) implies

$$\frac{u_j}{2}\left(\mathrm{R}_\ell(h_{D_{u_j}}) - \mathrm{R}_\ell(h^*)\right) < \frac{u_j}{2}\left(\mathrm{R}_\ell(h_{D_{u_j}}; \mathcal{L}_{u_j}) - \mathrm{R}_\ell(h^*; \mathcal{L}_{u_j}) + \hat{U}_{u_j}\right).$$

Since we have already established that $h^* \in V^{(u_j)}$, (12.45) and (12.46) imply

$$\frac{u_j}{2}\left(\mathrm{R}_\ell(h_{D_{u_j}}; \mathcal{L}_{u_j}) - \mathrm{R}_\ell(h^*; \mathcal{L}_{u_j}) + \hat{U}_{u_j}\right)$$
$$= (|Q_{u_j}| \vee 1)\left(\mathrm{R}_\ell(h; Q_{u_j}) - \mathrm{R}_\ell(h^*; Q_{u_j}) + \hat{U}_\ell(V^{(u_j)}; Q_{u_j}, \hat{\mathfrak{s}}(u_j))\right).$$

The definition of $\tilde{V}^{(u_j)}$ from Step 6 implies

$$(|Q_{u_j}| \vee 1)\left(\mathrm{R}_\ell(h; Q_{u_j}) - \mathrm{R}_\ell(h^*; Q_{u_j}) + \hat{U}_\ell(V^{(u_j)}; Q_{u_j}, \hat{\mathfrak{s}}(u_j))\right)$$
$$\leq (|Q_{u_j}| \vee 1)\left(2\hat{U}_\ell(V^{(u_j)}; Q_{u_j}, \hat{\mathfrak{s}}(u_j))\right).$$

By (12.46) and (12.49),

$$(|Q_{u_j}| \vee 1)\left(2\hat{U}_\ell(V^{(u_j)}; Q_{u_j}, \hat{\mathfrak{s}}(u_j))\right) = u_j\hat{U}_{u_j} < u_j\tilde{U}_\ell\left(V_{D_{u_j}}^{(u_j)}; \mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right).$$

Altogether, we have that, $\forall h \in \tilde{V}^{(u_j)}$,

$$\mathrm{R}_\ell(h_{D_{u_j}}) - \mathrm{R}_\ell(h^*) < 2\tilde{U}_\ell\left(V_{D_{u_j}}^{(u_j)}; \mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right). \tag{12.50}$$

By definition of $\mathring{\mathrm{M}}_\ell$, monotonicity of $m \mapsto \mathring{U}_\ell(\cdot, \cdot; \cdot, m, \cdot)$, and the condition on u_j in (12.12), we know that

$$\mathring{U}_\ell\left(\mathcal{F}_j, 2^{2-j}; \mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right) \leq 2^{-j-1}.$$

The fact that $u_j \geq 2u_{j-2}$, combined with the inductive hypothesis, implies

$$V^{(u_j)} \subseteq \tilde{V}^{(u_{j-2})} \subseteq \mathcal{F}\left(\mathcal{E}_\ell(2^{2-j}); {}_{01}\right).$$

This also implies $D_{u_j} \subseteq \mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{2-j}); {}_{01}))$. Combined with (12.7), these imply

$$\mathring{U}_\ell\left(V_{D_{u_j}}^{(u_j)}, 2^{2-j}; \mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right) \leq 2^{-j-1}.$$

Together with (12.6), this implies

$$\tilde{U}_\ell\left(V_{D_{u_j}}^{(u_j)}(2^{2-j};\ell);\mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right) \leq 2^{-j-1}.$$

The inductive hypothesis implies $V_{D_{u_j}}^{(u_j)} = V_{D_{u_j}}^{(u_j)}(2^{2-j};\ell)$, which means

$$\tilde{U}_\ell\left(V_{D_{u_j}}^{(u_j)};\mathcal{P}_{XY}, u_j/2, \hat{\mathfrak{s}}(u_j)\right) \leq 2^{-j-1}.$$

Plugging this into (12.50) implies, $\forall h \in \tilde{V}^{(u_j)}$,

$$\mathrm{R}_\ell(h_{D_{u_j}}) - \mathrm{R}_\ell(h^*) < 2^{-j}. \tag{12.51}$$

In particular, since $h^* \in \mathcal{F}$, we always have $\tilde{V}_{D_{u_j}}^{(u_j)} \subseteq [\mathcal{F}]$, so that (12.51) establishes that $\tilde{V}_{D_{u_j}}^{(u_j)} \subseteq [\mathcal{F}](2^{-j};\ell)$. Furthermore, since $h^* \in V^{(u_j)}$ on E, $\mathrm{sign}(h_{D_{u_j}}) = \mathrm{sign}(h)$ for every $h \in \tilde{V}^{(u_j)}$, so that every $h \in \tilde{V}^{(u_j)}$ has $\mathrm{er}(h) = \mathrm{er}(h_{D_{u_j}})$, and therefore (by definition of $\mathcal{E}_\ell(\cdot)$), (12.51) implies

$$\mathrm{er}(h) - \mathrm{er}(h^*) = \mathrm{er}(h_{D_{u_j}}) - \mathrm{er}(h^*) \leq \mathcal{E}_\ell\left(2^{-j}\right).$$

This implies $\tilde{V}^{(u_j)} \subseteq \mathcal{F}\left(\mathcal{E}_\ell(2^{-j});_{01}\right)$, which completes the inductive proof. This implies that, on the event E, if $u_{j_\varepsilon} \in S$, then (by monotonicity of $\mathcal{E}_\ell(\cdot)$ and the fact that $\mathcal{E}_\ell(\Gamma_\ell(\varepsilon)) \leq \varepsilon$)

$$\tilde{V}^{(u_{j_\varepsilon})} \subseteq \mathcal{F}(\mathcal{E}_\ell(2^{-j_\varepsilon});_{01}) \subseteq \mathcal{F}(\mathcal{E}_\ell(\Gamma_\ell(\varepsilon));_{01}) \subseteq \mathcal{F}(\varepsilon;_{01}).$$

In particular, since the update in Step 6 always keeps at least one element in V, the function $\hat{h}$ in Step 8 exists, and has $\hat{h} \in \tilde{V}^{(u_{j_\varepsilon})}$ (if $u_{j_\varepsilon} \in S$). Thus, on the event E, if $u_{j_\varepsilon} \in S$, then $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$. Therefore, since $u \geq u_{j_\varepsilon}$, to complete the proof it suffices to show that taking n of the size indicated in the theorem statement suffices to guarantee $u_{j_\varepsilon} \in S$, on an event (which includes E) having at least the stated probability.

Note that for any $j \in \{j_\ell, \ldots, j_\varepsilon\}$ with $u_{j-1} \in S \cup \{1\}$, every $m \in \{u_{j-1}+1, \ldots, u_j\} \cap S$ has $V^{(m)} \subseteq \tilde{V}^{(u_{j-1})}$; furthermore, we showed above that on the event E, if $u_{j-1} \in S$, then $\tilde{V}^{(u_{j-1})} \subseteq \mathcal{F}(\mathcal{E}_\ell(2^{1-j});_{01})$, so that $\mathrm{DIS}(V^{(m)}) \subseteq \mathrm{DIS}(\tilde{V}^{(u_{j-1})}) \subseteq \mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{1-j});_{01})) \subseteq \mathcal{U}_j$. Thus, on the event E, to guarantee $u_{j_\varepsilon} \in S$, it suffices to have

$$n \geq \sum_{j=j_\ell}^{j_\varepsilon} \sum_{m=u_{j-1}+1}^{u_j} \mathbb{I}_{\mathcal{U}_j}(X_m).$$

266

Noting that this is a sum of independent Bernoulli random variables, a Chernoff bound implies that on an event E' of probability at least $1 - 2^{-s}$,

$$\sum_{j=j_\ell}^{j_\varepsilon} \sum_{m=u_{j-1}+1}^{u_j} \mathbb{I}_{\mathcal{U}_j}(X_m) \leq s + 2e \sum_{j=j_\ell}^{j_\varepsilon} \sum_{m=u_{j-1}+1}^{u_j} \mathcal{P}(\mathcal{U}_j)$$

$$= s + 2e \sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j)(u_j - u_{j-1}) \leq s + 2e \sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j) u_j.$$

Thus, for n satisfying the condition in the theorem statement, on the event $E \cap E'$, we have $u_{j_\varepsilon} \in S$, and therefore (as proven above) $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$. Finally, a union bound implies that the event $E \cap E'$ has probability at least

$$1 - 2^{-s} - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{s}(2^i)},$$

as required. $\qquad\square$

Proof of Lemma 12.8. If $P\left(\overline{\mathrm{DISF}}(\mathcal{H})\right) = 0$, then $\phi_\ell(\mathcal{H}; m, P) = 0$, so that in this case, $\mathring{\phi}'_\ell$ trivially satisfies (12.5). Otherwise, suppose $P\left(\overline{\mathrm{DISF}}(\mathcal{H})\right) > 0$. By the classic symmetrization inequality [e.g., van der Vaart and Wellner, 1996, Lemma 2.3.1],

$$\phi_\ell(\mathcal{H}, m, P) \leq 2\mathbb{E}\left[\left|\hat{\phi}_\ell(\mathcal{H}; Q, \Xi_{[m]})\right|\right],$$

where $Q \sim P^m$ and $\Xi_{[m]} = \{\xi_1, \ldots, \xi_m\} \sim \mathrm{Uniform}(\{-1, +1\}^m)$ are independent. Fix any measurable $\mathcal{U} \supseteq \overline{\mathrm{DISF}}(\mathcal{H})$. Then

$$\mathbb{E}\left[\left|\hat{\phi}_\ell(\mathcal{H}; Q, \Xi_{[m]})\right|\right] = \mathbb{E}\left[\left|\hat{\phi}_\ell(\mathcal{H}; Q \cap \mathcal{U}, \Xi_{[|Q \cap \mathcal{U}|]})\right| \frac{|Q \cap \mathcal{U}|}{m}\right], \tag{12.52}$$

where $\Xi_{[q]} = \{\xi_1, \ldots, \xi_q\}$ for any $q \in \{0, \ldots, m\}$. By the classic desymmetrization inequality [see e.g., Koltchinskii, 2008], applied under the conditional distribution given $|Q \cap \mathcal{U}|$, the right hand side of (12.52) is at most

$$\mathbb{E}\left[2\phi_\ell(\mathcal{H}, |Q \cap \mathcal{U}|, P_\mathcal{U}) \frac{|Q \cap \mathcal{U}|}{m}\right] + \sup_{h, g \in \mathcal{H}} |\mathrm{R}_\ell(h; P_\mathcal{U}) - \mathrm{R}_\ell(g; P_\mathcal{U})| \frac{\mathbb{E}\left[\sqrt{|Q \cap \mathcal{U}|}\right]}{m}. \tag{12.53}$$

By Jensen's inequality, the second term in (12.53) is at most

$$\sup_{h,g\in\mathcal{H}} |R_\ell(h;P_\mathcal{U}) - R_\ell(g;P_\mathcal{U})|\sqrt{\frac{P(\mathcal{U})}{m}} \leq D_\ell(\mathcal{H};P_\mathcal{U})\sqrt{\frac{P(\mathcal{U})}{m}} = D_\ell(\mathcal{H};P)\sqrt{\frac{1}{m}}.$$

Decomposing based on $|Q\cap\mathcal{U}|$, the first term in (12.53) is at most

$$\mathbb{E}\left[2\phi_\ell(\mathcal{H}, |Q\cap\mathcal{U}|, P_\mathcal{U})\frac{|Q\cap\mathcal{U}|}{m}\mathbb{I}\left[|Q\cap\mathcal{U}| \geq (1/2)P(\mathcal{U})m\right]\right]$$

$$+ 2\bar{\ell}P(\mathcal{U})\mathbb{P}\left(|Q\cap\mathcal{U}| < (1/2)P(\mathcal{U})m\right). \quad (12.54)$$

Since $|Q\cap\mathcal{U}| \geq (1/2)P(\mathcal{U})m \Rightarrow |Q\cap\mathcal{U}| \geq \lceil(1/2)P(\mathcal{U})m\rceil$, and $\phi_\ell(\mathcal{H}, q, P_\mathcal{U})$ is nonincreasing in q, the first term in (12.54) is at most

$$2\phi_\ell(\mathcal{H}, \lceil(1/2)P(\mathcal{U})m\rceil, P_\mathcal{U})\mathbb{E}\left[\frac{|Q\cap\mathcal{U}|}{m}\right] = 2\phi_\ell(\mathcal{H}, \lceil(1/2)P(\mathcal{U})m\rceil, P_\mathcal{U})P(\mathcal{U}),$$

while a Chernoff bound implies the second term in (12.54) is at most

$$2\bar{\ell}P(\mathcal{U})\exp\{-P(\mathcal{U})m/8\} \leq \frac{16\bar{\ell}}{m}.$$

Plugging back into (12.53), we have

$$\phi_\ell(\mathcal{H}, m, P) \leq 4\phi_\ell(\mathcal{H}, \lceil(1/2)P(\mathcal{U})m\rceil, P_\mathcal{U})P(\mathcal{U}) + \frac{32\bar{\ell}}{m} + 2D_\ell(\mathcal{H};P)\sqrt{\frac{1}{m}}. \quad (12.55)$$

Next, note that, for any $\sigma \geq D_\ell(\mathcal{H};P)$, $\frac{\sigma}{\sqrt{P(\mathcal{U})}} \geq D_\ell(\mathcal{H};P_\mathcal{U})$. Also, if $\mathcal{U} = \mathcal{U}' \times \mathcal{Y}$ for some $\mathcal{U}' \supseteq \mathrm{DISF}(\mathcal{H})$, then $h^*_{P_\mathcal{U}} = h^*_P$, so that if $h^*_P \in \mathcal{H}$, (12.5) implies

$$\phi_\ell(\mathcal{H}, \lceil(1/2)P(\mathcal{U})m\rceil, P_\mathcal{U}) \leq \mathring{\phi}_\ell\left(\frac{\sigma}{\sqrt{P(\mathcal{U})}}, \mathcal{H}; \lceil(1/2)P(\mathcal{U})m\rceil, P_\mathcal{U}\right). \quad (12.56)$$

Combining (12.55) with (12.56), we see that $\mathring{\phi}'_\ell$ satisfies the condition (12.5) of Definition 12.5.

Furthermore, by the fact that $\mathring{\phi}_\ell$ satisfies (12.4) of Definition 12.5, combined with the monotonicity imposed by the infimum in the definition of $\mathring{\phi}'_\ell$, it is easy to check that $\mathring{\phi}'_\ell$ also satisfies (12.4) of Definition 12.5. In particular, note that any $\mathcal{H}'' \subseteq \mathcal{H}' \subseteq [\mathcal{F}]$ and $\mathcal{U}'' \subseteq \mathcal{X}$ have $\mathrm{DISF}(\mathcal{H}''_{\mathcal{U}''}) \subseteq \mathrm{DISF}(\mathcal{H}')$, so that the range of $\mathcal{U}$ in the infimum is never smaller for $\mathcal{H} = \mathcal{H}''_{\mathcal{U}''}$ relative to that for $\mathcal{H} = \mathcal{H}'$. $\square$

Proof of Corollary 12.9. Let $\mathring{\phi}'_\ell$ be as in Lemma 12.8, and define for any $m \in \mathbb{N}$, $s \in [1, \infty)$, $\zeta \in [0, \infty]$, and $\mathcal{H} \subseteq [\mathcal{F}]$,

$$
\begin{aligned}
&\mathring{U}'_\ell(\mathcal{H}, \zeta; \mathcal{P}_{XY}, m, s) \\
&= \tilde{K}\left(\mathring{\phi}'_\ell(\mathrm{D}_\ell([\mathcal{H}](\zeta; \ell)), \mathcal{H}; m, \mathcal{P}_{XY}) + \mathrm{D}_\ell([\mathcal{H}](\zeta; \ell)) \sqrt{\frac{s}{m}} + \frac{\bar{\ell}s}{m} \right).
\end{aligned}
$$

That is, $\mathring{U}'_\ell$ is the function $\mathring{U}_\ell$ that would result from using $\mathring{\phi}'_\ell$ in place of $\mathring{\phi}_\ell$. Let $\mathcal{U} = \mathrm{DISF}(\mathcal{H})$, and suppose $\mathcal{P}(\mathcal{U}) > 0$. Then since $\mathrm{DISF}([\mathcal{H}]) = \mathrm{DISF}(\mathcal{H})$ implies

$$
\begin{aligned}
\mathrm{D}_\ell([\mathcal{H}](\zeta; \ell)) &= \mathrm{D}_\ell([\mathcal{H}](\zeta; \ell); \mathcal{P}_{\mathcal{U}}) \sqrt{\mathcal{P}(\mathcal{U})} \\
&= \mathrm{D}_\ell([\mathcal{H}](\zeta/\mathcal{P}(\mathcal{U}); \ell, \mathcal{P}_{\mathcal{U}}); \mathcal{P}_{\mathcal{U}}) \sqrt{\mathcal{P}(\mathcal{U})},
\end{aligned}
$$

a little algebra reveals that for $m \geq 2\mathcal{P}(\mathcal{U})^{-1}$,

$$
\mathring{U}'_\ell(\mathcal{H}, \zeta; \mathcal{P}_{XY}, m, s) \leq 33\mathcal{P}(\mathcal{U})\mathring{U}_\ell(\mathcal{H}, \zeta/\mathcal{P}(\mathcal{U}); \mathcal{P}_{\mathcal{U}}, \lceil (1/2)\mathcal{P}(\mathcal{U})m \rceil, s). \tag{12.57}
$$

In particular, for $j \geq j_\ell$, taking $\mathcal{H} = \mathcal{F}_j$, we have (from the definition of $\mathcal{F}_j$) $\mathcal{U} = \mathrm{DISF}(\mathcal{H}) = \mathrm{DIS}(\mathcal{H}) = \mathcal{U}_j$, so that when $\mathcal{P}(\mathcal{U}_j) > 0$, any

$$
m \geq 2\mathcal{P}(\mathcal{U}_j)^{-1}\mathring{\mathrm{M}}_\ell\left(\frac{2^{-j-1}}{33\mathcal{P}(\mathcal{U}_j)}, \frac{2^{2-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, \hat{\mathfrak{s}}(2m) \right)
$$

suffices to make the right side of (12.57) (with $s = \hat{\mathfrak{s}}(2m)$ and $\zeta = 2^{2-j}$) at most 2^{-j-1}; in particular, this means taking u_j equal to $2m \vee u_{j-1} \vee 2u_{j-2}$ for any such m (with $\log_2(m) \in \mathbb{N}$) suffices to satisfy (12.12) (with the $\mathring{\mathrm{M}}_\ell$ in (12.12) defined with respect to the $\mathring{\phi}'_\ell$ function); monotonicity of $\zeta \mapsto \mathring{\mathrm{M}}_\ell\left(\zeta, \frac{2^{2-j}}{\mathcal{P}(\mathcal{U}_j)}; \mathcal{F}_j, \mathcal{P}_{\mathcal{U}_j}, \hat{\mathfrak{s}}(2m) \right)$ implies (12.14) is a sufficient condition for this. In the special case where $\mathcal{P}(\mathcal{U}_j) = 0$, $\mathring{U}'_\ell(\mathcal{F}_j, 2^{2-j}; \mathcal{P}_{XY}, m, s) = \tilde{K}\frac{\bar{\ell}s}{m}$, so that taking $u_j \geq \tilde{K}\bar{\ell}\hat{\mathfrak{s}}(u_j)2^{j+2} \vee u_{j-1} \vee 2u_{j-1}$ suffices to satisfy (12.12) (again, with the $\mathring{\mathrm{M}}_\ell$ in (12.12) defined in terms of $\mathring{\phi}'_\ell$). Plugging these values into Theorem 12.7 completes the proof. $\qquad\square$

Proof of Theorem 12.16. Let $\tilde{j}_\varepsilon = \lceil \log_2(1/\Psi_\ell(\varepsilon)) \rceil$. For $j_\ell \leq j \leq \tilde{j}_\varepsilon$, let $s_j = \mathrm{Log}\left(\frac{48\left(2+\tilde{j}_\varepsilon-j\right)^2}{\delta} \right)$, and define $u_j = 2^{\lceil \log_2(u'_j) \rceil}$, where

$$
u'_j = c'\left(b2^{j(2-\beta)} + \bar{\ell}2^j \right)\left(\mathrm{vc}\left(\mathcal{G}_\mathcal{F} \right)\mathrm{Log}\left(\chi_\ell \bar{\ell} \right) + s_j \right), \tag{12.58}
$$

for an appropriate universal constant $c' \in [1, \infty)$. A bit of calculus reveals that for $j_\ell + 2 \leq j \leq \tilde{j}_\varepsilon$, $u'_j \geq u'_{j-1}$ and $u'_j \geq 2u'_{j-2}$, so that $u_j \geq u_{j-1}$ and $u_j \geq 2u_{j-2}$ as well; this is also trivially satisfied for $j \in \{j_\ell, j_\ell + 1\}$ if we take $u_{j-2} = 1$ in these cases (as in Theorem 12.7). Combining this fact with (12.31), (12.8), and (12.9), we find that, for an appropriate choice of the constant c', these u_j satisfy (12.12) when we define $\hat{\mathfrak{s}}$ such that, for every $j \in \{j_\ell, \ldots, \tilde{j}_\varepsilon\}$, $\forall m \in \{2u_{j-1}, \ldots, u_j\}$ with $\log_2(m) \in \mathbb{N}$,

$$\hat{\mathfrak{s}}(m) = \text{Log}\left(\frac{12 \log_2\left(4u_j/m\right)^2 \left(2 + \tilde{j}_\varepsilon - j\right)^2}{\delta}\right).$$

Additionally, let $s = \log_2(2/\delta)$.

Next, note that, since $\Psi_\ell(\varepsilon) \leq \Gamma_\ell(\varepsilon)$ and u_j is nondecreasing in j,

$$u_{j_\varepsilon} \leq u_{\tilde{j}_\varepsilon} \leq 26c'\left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)\left(\text{vc}\left(\mathcal{G}_\mathcal{F}\right)\text{Log}\left(\chi_\ell\bar{\ell}\right) + \text{Log}(1/\delta)\right),$$

so that, for any $c \geq 26c'$, we have $u \geq u_{i_\varepsilon}$, as required by Theorem 12.7.

For $\mathcal{U}_j$ as in Theorem 12.7, note that by Condition 12.10 and the definition of θ,

$$\mathcal{P}\left(\mathcal{U}_j\right) = \mathcal{P}\left(\text{DIS}\left(\mathcal{F}\left(\mathcal{E}_\ell\left(2^{2-j}\right); {}_{01}\right)\right)\right) \leq \mathcal{P}\left(\text{DIS}\left(\text{B}\left(h^*, a\mathcal{E}_\ell\left(2^{2-j}\right)^\alpha\right)\right)\right)$$

$$\leq \theta \max\left\{a\mathcal{E}_\ell\left(2^{2-j}\right)^\alpha, a\varepsilon^\alpha\right\} \leq \theta \max\left\{a\Psi_\ell^{-1}\left(2^{2-j}\right)^\alpha, a\varepsilon^\alpha\right\}.$$

Because Ψ_ℓ is strictly increasing on $(0, 1)$, for $j \leq \tilde{j}_\varepsilon$, $\Psi_\ell^{-1}\left(2^{2-j}\right) \geq \varepsilon$, so that this last expression is equal to $\theta a \Psi_\ell^{-1}\left(2^{2-j}\right)^\alpha$. This implies

$$\sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}\left(\mathcal{U}_j\right)u_j \leq \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \mathcal{P}\left(\mathcal{U}_j\right)u_j$$

$$\lesssim \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} a\theta\Psi_\ell^{-1}\left(2^{2-j}\right)^\alpha\left(b2^{j(2-\beta)} + \bar{\ell}2^j\right)\left(A_1 + \text{Log}\left(2 + \tilde{j}_\varepsilon - j\right)\right). \tag{12.59}$$

We can change the order of summation in the above expression by letting $i = \tilde{j}_\varepsilon - j$ and summing from 0 to $N = j_\varepsilon - j_\ell$. In particular, since $2^{\tilde{j}_\varepsilon} \leq 2/\Psi_\ell(\varepsilon)$, (12.59) is at most

$$\sum_{i=0}^{N} a\theta\Psi_\ell^{-1}\left(2^{2-\tilde{j}_\varepsilon}2^i\right)^\alpha\left(\frac{4b2^{i(\beta-2)}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{2\bar{\ell}2^{-i}}{\Psi_\ell(\varepsilon)}\right)\left(A_1 + \text{Log}(i+2)\right). \tag{12.60}$$

270

Since $x \mapsto \Psi_\ell^{-1}(x)/x$ is nonincreasing on $(0, \infty)$, $\Psi_\ell^{-1}\left(2^{2-\tilde{j}_\varepsilon}2^i\right) \leq 2^{i+2}\Psi_\ell^{-1}\left(2^{-\tilde{j}_\varepsilon}\right)$, and since Ψ_ℓ^{-1} is increasing, this latter expression is at most $2^{i+2}\Psi_\ell^{-1}\left(\Psi_\ell(\varepsilon)\right) = 2^{i+2}\varepsilon$. Thus, (12.60) is at most

$$16 a \theta \varepsilon^\alpha \sum_{i=0}^{N} \left(\frac{b 2^{i(\alpha+\beta-2)}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell} 2^{i(\alpha-1)}}{\Psi_\ell(\varepsilon)}\right) (A_1 + \mathrm{Log}(i+2)). \tag{12.61}$$

In general, $\mathrm{Log}(i+2) \leq \mathrm{Log}(N+2)$, so that $\sum_{i=0}^{N} 2^{i(\alpha+\beta-2)} (A_1 + \mathrm{Log}(i+2)) \leq (A_1 + \mathrm{Log}(N+2))(N+1)$ and $\sum_{i=0}^{N} 2^{i(\alpha-1)} (A_1 + \mathrm{Log}(i+2)) \leq (A_1 + \mathrm{Log}(N+2))(N+1)$. When $\alpha + \beta < 2$, we also have $\sum_{i=0}^{N} 2^{i(\alpha+\beta-2)} \leq \sum_{i=0}^{\infty} 2^{i(\alpha+\beta-2)} = \frac{1}{1-2^{(\alpha+\beta-2)}}$ and $\sum_{i=0}^{N} 2^{i(\alpha+\beta-2)}\mathrm{Log}(i+2) \leq \sum_{i=0}^{\infty} 2^{i(\alpha+\beta-2)}\mathrm{Log}(i+2) \leq \frac{2}{1-2^{(\alpha+\beta-2)}}\mathrm{Log}\left(\frac{1}{1-2^{(\alpha+\beta-2)}}\right)$. Similarly, if $\alpha < 1$, $\sum_{i=0}^{N} 2^{i(\alpha-1)} \leq \sum_{i=0}^{\infty} 2^{i(\alpha-1)} = \frac{1}{1-2^{(\alpha-1)}}$ and likewise $\sum_{i=0}^{N} 2^{i(\alpha-1)}\mathrm{Log}(i+2) \leq \sum_{i=0}^{\infty} 2^{i(\alpha-1)}\mathrm{Log}(i+2) \leq \frac{2}{1-2^{(\alpha-1)}}\mathrm{Log}\left(\frac{1}{1-2^{(\alpha-1)}}\right)$. By combining these observations (along with a convention that $\frac{1}{1-2^{(\alpha-1)}} = \infty$ when $\alpha = 1$, and $\frac{1}{1-2^{(\alpha+\beta-2)}} = \infty$ when $\alpha = \beta = 1$), we find that (12.61) is

$$\lesssim a \theta \varepsilon^\alpha \left(\frac{b(A_1 + \mathrm{Log}(B_1))B_1}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}(A_1 + \mathrm{Log}(C_1))C_1}{\Psi_\ell(\varepsilon)}\right).$$

Thus, for an appropriately large numerical constant c, any n satisfying (12.33) has

$$n \geq s + 2e \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \mathcal{P}(\mathcal{U}_j)u_j,$$

as required by Theorem 12.7.

Finally, we need to show the success probability from Theorem 12.7 is at least $1 - \delta$, for $\hat{\mathfrak{s}}$ and s as above. Toward this end, note that

$$\sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)}$$

$$\leq \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \sum_{i=\log_2(u_{j-1})+1}^{\log_2(u_j)} \frac{\delta}{2\left(2 + \log_2(u_j) - i\right)^2 \left(2 + \tilde{j}_\varepsilon - j\right)^2}$$

$$= \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \sum_{t=0}^{\log_2(u_j/u_{j-1})-1} \frac{\delta}{2(2+t)^2 \left(2 + \tilde{j}_\varepsilon - j\right)^2}$$

$$< \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \frac{\delta}{2\left(2 + \tilde{j}_\varepsilon - j\right)^2} < \sum_{t=0}^{\infty} \frac{\delta}{2(2+t)^2} < \delta/2.$$

271

Noting that $2^{-s} = \delta/2$, we find that indeed

$$1 - 2^{-s} - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)} \geq 1 - \delta.$$

Therefore, Theorem 12.7 implies the stated result. $\qquad\square$

Proof Sketch of Theorem 12.17. The proof follows analogously to that of Theorem 12.16, with the exception that now, for each integer j with $j_\ell \leq j \leq \tilde{j}_\varepsilon$, we replace the definition of u'_j from (12.58) with the following definition. Letting $c_j = \text{vc}(\mathcal{G}_\mathcal{F})\text{Log}\left(\left(\bar{\ell}/b\right)\left(a\theta 2^j \Psi_\ell^{-1}(2^{2-j})^\alpha\right)^\beta\right)$, define

$$u'_j = c'\left(b2^{j(2-\beta)}\left(a\theta\Psi_\ell^{-1}(2^{2-j})^\alpha\right)^{1-\beta} + \bar{\ell}2^j\right)(c_j + s_j),$$

where $c' \in [1, \infty)$ is an appropriate universal constant, and s_j is as in the proof of Theorem 12.16. With this substitution in place, the values u_j and s, and function $\hat{\mathfrak{s}}$, are then defined as in the proof of Theorem 12.16. Since $x \mapsto x\Psi_\ell^{-1}(1/x)$ is nondecreasing, a bit of calculus reveals $u_j \geq u_{j-1}$ and $u_j \geq 2u_{j-2}$. Combined with (12.35), (12.9), (12.8), and Lemma 12.12, this implies we can choose the constant c' so that these u_j satisfy (12.14). By an identical argument to that used in Theorem 12.16, we have

$$1 - 2^{-s} - \sum_{i=1}^{\log_2(u_{j_\varepsilon})} 6e^{-\hat{\mathfrak{s}}(2^i)} \geq 1 - \delta.$$

It remains only to show that any values of u and n satisfying (12.36) and (12.37), respectively, necessarily also satisfy the respective conditions for u and n in Corollary 12.9.

Toward this end, note that since $x \mapsto x\Psi_\ell^{-1}(1/x)$ is nondecreasing on $(0, \infty)$, we have that

$$u_{j_\varepsilon} \leq u_{\tilde{j}_\varepsilon} \lesssim \left(\frac{b\left(a\theta\varepsilon^\alpha\right)^{1-\beta}}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)}\right)A_2.$$

Thus, for an appropriate choice of c, any u satisfying (12.36) has $u \geq u_{j_\varepsilon}$, as required by Corollary 12.9.

272

Finally, note that for $\mathcal{U}_j$ as in Theorem 12.7, and $i_j = \tilde{j}_\varepsilon - j$,

$$\sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j)u_j \leq \sum_{j=j_\ell}^{j_\varepsilon} a\theta\Psi_\ell^{-1}(2^{2-j})^\alpha u_j$$

$$\lesssim \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} b\left(a\theta 2^j \Psi_\ell^{-1}(2^{2-j})^\alpha\right)^{2-\beta}(A_2 + \mathrm{Log}\,(i_j + 2))$$

$$+ \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \bar{\ell}a\theta 2^j \Psi_\ell^{-1}(2^{2-j})^\alpha (A_2 + \mathrm{Log}\,(i_j + 2)).$$

By changing the order of summation, now summing over values of i_j from 0 to $N = \tilde{j}_\varepsilon - j_\ell \leq \log_2(4\bar{\ell}/\Psi_\ell(\varepsilon))$, and noting $2^{\tilde{j}_\varepsilon} \leq 2/\Psi_\ell(\varepsilon)$, and $\Psi_\ell^{-1}(2^{-\tilde{j}_\varepsilon}2^{2+i}) \leq 2^{2+i}\varepsilon$ for $i \geq 0$, this last expression is

$$\lesssim \sum_{i=0}^{N} b\left(\frac{a\theta 2^{i(\alpha-1)}\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}\right)^{2-\beta}(A_2 + \mathrm{Log}\,(i+2)) \tag{12.62}$$

$$+ \sum_{i=0}^{N} \frac{\bar{\ell}a\theta 2^{i(\alpha-1)}\varepsilon^\alpha}{\Psi_\ell(\varepsilon)}(A_2 + \mathrm{Log}\,(i+2)).$$

Considering these sums separately, we have $\sum_{i=0}^{N} 2^{i(\alpha-1)(2-\beta)}(A_2 + \mathrm{Log}(i+2)) \leq (N+1)(A_2 + \mathrm{Log}(N+2))$ and $\sum_{i=0}^{N} 2^{i(\alpha-1)}(A_2 + \mathrm{Log}(i+2)) \leq (N+1)(A_2 + \mathrm{Log}(N+2))$. When $\alpha < 1$, we also have $\sum_{i=0}^{N} 2^{i(\alpha-1)(2-\beta)}(A_2 + \mathrm{Log}(i+2)) \leq \sum_{i=0}^{\infty} 2^{i(\alpha-1)(2-\beta)}(A_2 + \mathrm{Log}(i+2)) \leq \frac{2}{1-2^{(\alpha-1)(2-\beta)}}\mathrm{Log}\left(\frac{1}{1-2^{(\alpha-1)(2-\beta)}}\right) + \frac{1}{1-2^{(\alpha-1)(2-\beta)}}A_2$, and similarly $\sum_{i=0}^{N} 2^{i(\alpha-1)}(A_2 + \mathrm{Log}(i+2)) \leq \frac{1}{1-2^{(\alpha-1)}}A_2 + \frac{2}{1-2^{(\alpha-1)}}\mathrm{Log}\left(\frac{1}{1-2^{(\alpha-1)}}\right)$. Thus, generally $\sum_{i=0}^{N} 2^{i(\alpha-1)(2-\beta)}(A_2 + \mathrm{Log}(i+2)) \lesssim B_2(A_2 + \mathrm{Log}(B_2))$ and $\sum_{i=0}^{N} 2^{i(\alpha-1)}(A_2 + \mathrm{Log}(i+2)) \lesssim C_2(A_2 + \mathrm{Log}(C_2))$. Plugging this into (12.62), we find that for an appropriately large numerical constant c, any n satisfying (12.37) has $n \geq \sum_{j=j_\ell}^{j_\varepsilon} \mathcal{P}(\mathcal{U}_j)u_j$, as required by Corollary 12.9. $\qquad\square$

12.7 Results for Efficiently Computable Updates

Here we include more detailed sketches of the arguments leading to computationally efficient variants of Algorithm 1, for which the specific results proven above for the given applications

273

remain valid. Throughout this section, we adopt the notational conventions introduced in the proof of Theorem 12.7 (e.g., $V^{(m)}$, $\tilde{V}^{(m)}$, Q_m, $\mathcal{L}_m$, S), except in each instance here these are defined in the context of applying Algorithm 1 with the respective stated variant of $\hat{T}_\ell$.

12.7.1 Proof of Theorem 12.16 under (12.34)

We begin with the application to VC Subgraph classes, first showing that if we specify $\hat{T}_\ell(V; Q, m)$ as in (12.34), the conclusions of Theorem 12.16 remain valid. Fix any $\hat{\mathfrak{s}}$ function (to be specified below), and fix any value of $\varepsilon \in (0, 1)$. First note that, for any m with $\log_2(m) \in \mathbb{N}$, by a Chernoff bound and the law of total probability, on an event E''_m of probability at least $1 - 2^{1-\hat{\mathfrak{s}}(m)}$, if $m \in S$, then

$$(1/2)m\mathcal{P}(D_m) - \sqrt{\hat{\mathfrak{s}}(m)m\mathcal{P}(D_m)} \leq |Q_m| \leq \hat{\mathfrak{s}}(m) + em\mathcal{P}(D_m). \tag{12.63}$$

Also recall that, for any m with $\log_2(m) \in \mathbb{N}$, by Lemma 12.4 and the law of total probability, on an event E_m of probability at least $1 - 6e^{-\hat{\mathfrak{s}}(m)}$, if $m \in S$ and $h^* \in V^{(m)}$, then

$$(|Q_m| \vee 1)\left(\mathrm{R}_\ell(h^*; Q_m) - \inf_{g \in V^{(m)}} \mathrm{R}_\ell(g; Q_m)\right)$$
$$= \frac{m}{2}\left(\mathrm{R}_\ell(h^*; \mathcal{L}_m) - \inf_{g_{D_m} \in V^{(m)}_{D_m}} \mathrm{R}_\ell(g_{D_m}; \mathcal{L}_m)\right)$$
$$< \frac{m}{2}\tilde{U}_\ell\left(V^{(m)}_{D_m}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m)\right) \tag{12.64}$$

and $\forall h \in \tilde{V}^{(m)}$,

$$\frac{m}{2}\left(\mathrm{R}_\ell(h_{D_m}) - \mathrm{R}_\ell(h^*)\right)$$
$$< \frac{m}{2}\left(\mathrm{R}_\ell(h_{D_m}; \mathcal{L}_m) - \mathrm{R}_\ell(h^*; \mathcal{L}_m) + \tilde{U}_\ell\left(V^{(m)}_{D_m}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m)\right) \wedge \bar{\ell}\right)$$
$$= |Q_m|\left(\mathrm{R}_\ell(h; Q_m) - \mathrm{R}_\ell(h^*; Q_m)\right) + \frac{m}{2}\left(\tilde{U}_\ell\left(V^{(m)}_{D_m}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m)\right) \wedge \bar{\ell}\right)$$
$$\leq (|Q_m| \vee 1)\hat{T}_\ell\left(V^{(m)}; Q_m, m\right) + \frac{m}{2}\left(\tilde{U}_\ell\left(V^{(m)}_{D_m}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m)\right) \wedge \bar{\ell}\right). \tag{12.65}$$

Fix a value $i_\varepsilon \in \mathbb{N}$ (an appropriate value for which will be determined below), and let $\chi_\ell = \chi_\ell(\Psi_\ell(\varepsilon))$. For $m \in \mathbb{N}$ with $\log_2(m) \in \mathbb{N}$, let

$$\tilde{T}_\ell(m) = c_2 \left(\frac{b}{m} \left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}(\chi_\ell \bar{\ell}) + \hat{\mathfrak{s}}(m) \right) \right)^{\frac{1}{2-\beta}}$$

$$+ c_2 \frac{\bar{\ell}}{m} \left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}(\chi_\ell \bar{\ell}) + \hat{\mathfrak{s}}(m) \right),$$

for an appropriate universal constant $c_2 \in [1, \infty)$ (to be determined below); for completeness, also define $\tilde{T}_\ell(1) = \bar{\ell}$. We will now prove by induction that, for an appropriate value of the constant c_0 in (12.34), for any m' with $\log_2(m') \in \{1, \ldots, i_\varepsilon\}$, on the event $\bigcap_{i=1}^{\log_2(m')-1} E_{2^i} \cap E''_{2^{i+1}}$, if $m' \in S$, then $h^* \in V^{(m')}$,

$$V_{D_{m'}}^{(m')} \subseteq [\mathcal{F}](\hat{\gamma}_{m'/2}; \ell) \subseteq [\mathcal{F}](2\tilde{T}_\ell(m'/2) \vee \Psi_\ell(\varepsilon); \ell),$$

$$V^{(m')} \subseteq \mathcal{F}(\mathcal{E}_\ell(\hat{\gamma}_{m'/2}); \text{01}) \subseteq \mathcal{F}(\mathcal{E}_\ell(2\tilde{T}_\ell(m'/2) \vee \Psi_\ell(\varepsilon)); \text{01}),$$

$$\tilde{U}_\ell \left(V_{D_{m'}}^{(m')}; \mathcal{P}_{XY}, m'/2, \hat{\mathfrak{s}}(m') \right) \wedge \bar{\ell} \leq \frac{|Q_{m'}| \vee 1}{m'/2} \left(\hat{T}_\ell \left(V^{(m')}; Q_{m'}, m' \right) \wedge \bar{\ell} \right),$$

and if $\hat{\gamma}_{m'/2} \geq \Psi_\ell(\varepsilon)$,

$$\frac{|Q_{m'}| \vee 1}{m'/2} \left(\hat{T}_\ell \left(V^{(m')}; Q_{m'}, m' \right) \wedge \bar{\ell} \right) \leq \tilde{T}_\ell(m').$$

As a base case for this inductive argument, we note that for $m' = 2$, we have (by definition) $\hat{\gamma}_{m'/2} = \bar{\ell}$, and furthermore (if $c_0 \wedge c_2 \geq 2$) $\hat{T}_\ell(V^{(2)}; Q_2, 2) \geq \bar{\ell}$ and $\tilde{T}_\ell(1) \geq \bar{\ell}$, so that the claimed inclusions and inequalities trivially hold. Now, for the inductive step, take as an inductive hypothesis that the claim is satisfied for $m' = m$ for some $m \in \mathbb{N}$ with $\log_2(m) \in \{1, \ldots, i_\varepsilon - 1\}$. Suppose the event $\bigcap_{i=1}^{\log_2(m)} E_{2^i} \cap E''_{2^{i+1}}$ occurs, and that $2m \in S$. By the inductive hypothesis, combined with (12.64) and the fact that $(|Q_m| \vee 1)\mathrm{R}_\ell(h^*; Q_m) \leq (m/2)\bar{\ell}$, we have

$$(|Q_m| \vee 1) \left(\mathrm{R}_\ell(h^*; Q_m) - \inf_{g \in V^{(m)}} \mathrm{R}_\ell(g; Q_m) \right)$$

$$\leq \frac{m}{2} \left(\tilde{U}_\ell \left(V_{D_m}^{(m)}; \mathcal{P}_{XY}, m/2, \hat{\mathfrak{s}}(m) \right) \wedge \bar{\ell} \right) \leq (|Q_m| \vee 1)\hat{T}_\ell \left(V^{(m)}; Q_m, m \right).$$

275

Therefore, $h^* \in \tilde{V}^{(m)}$ as well, which implies $h^* \in V^{(2m)} = \tilde{V}^{(m)}$. Furthermore, by (12.65), the inductive hypothesis, and the definition of $\tilde{V}^{(m)}$ from Step 6, $\forall h \in V^{(2m)} = \tilde{V}^{(m)}$,

$$\mathrm{R}_\ell(h_{D_m}) - \mathrm{R}_\ell(h^*) < 2\frac{|Q_m| \vee 1}{m/2} \left(\hat{T}_\ell \left(V^{(m)}; Q_m, m \right) \wedge \bar{\ell} \right),$$

and if $\hat{\gamma}_{m/2} \geq \Psi_\ell(\varepsilon)$, then this is at most $2\tilde{T}_\ell(m)$.

Since $\hat{\gamma}_m = 2\frac{|Q_m| \vee 1}{m/2} \left(\hat{T}_\ell \left(V^{(m)}; Q_m, m \right) \wedge \bar{\ell} \right)$, and $\mathrm{R}_\ell(h_{D_{2m}}) \leq \mathrm{R}_\ell(h_{D_m})$ for every $h \in V^{(2m)}$d, we have $V^{(2m)}_{D_{2m}} \subseteq [\mathcal{F}](\hat{\gamma}_m; \ell) \subseteq [\mathcal{F}](2\tilde{T}_\ell(m) \vee \Psi_\ell(\varepsilon); \ell)$. By definition of $\mathcal{E}_\ell(\cdot)$, we also have $\mathrm{er}(h_{D_{2m}}) - \mathrm{er}(h^*) \leq \mathcal{E}_\ell(\hat{\gamma}_m)$ for every $h \in V^{(2m)}$; since $h^* \in V^{(2m)}$, we have $\mathrm{sign}(h_{D_{2m}}) = \mathrm{sign}(h)$, so that $\mathrm{er}(h) - \mathrm{er}(h^*) \leq \mathcal{E}_\ell(\hat{\gamma}_m)$ as well: that is, $V^{(2m)} \subseteq \mathcal{F}(\mathcal{E}_\ell(\hat{\gamma}_m); {}_{01}) \subseteq \mathcal{F}(\mathcal{E}_\ell(2\tilde{T}_\ell(m) \vee \Psi_\ell(\varepsilon)); {}_{01})$. Combining these facts with (12.5), (12.25), Condition 12.11, monotonicity of $\mathrm{vc}(\mathcal{G}_{\mathcal{H}_\mathcal{U}})$ in both $\mathcal{U}$ and $\mathcal{H}$, and the fact that $\|\mathrm{F}(\mathcal{G}_{V^{(2m)}_{D_{2m}}, \mathcal{P}_{XY}})\|^2_{\mathcal{P}_{XY}} \leq \bar{\ell}^2 \mathcal{P}(D_{2m})$, we have that

$$\tilde{U}_\ell \left(V^{(2m)}_{D_{2m}}; \mathcal{P}_{XY}, m, \hat{\mathfrak{s}}(2m) \right) \leq c_1 \sqrt{b\hat{\gamma}_m^\beta \frac{\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}\mathcal{P}(D_{2m})}{b\hat{\gamma}_m^\beta} \right) + \hat{\mathfrak{s}}(2m)}{m}}$$

$$+ c_1 \bar{\ell} \frac{\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}\mathcal{P}(D_{2m})}{b\hat{\gamma}_m^\beta} \right) + \hat{\mathfrak{s}}(2m)}{m}, \quad (12.66)$$

for some universal constant $c_1 \in [1, \infty)$. By (12.63), we have $\mathcal{P}(D_{2m}) \leq \frac{3}{m}(|Q_{2m}| + \hat{\mathfrak{s}}(2m))$, so that the right hand side of (12.66) is at most

$$c_1 \sqrt{b\hat{\gamma}_m^\beta \frac{\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}6(|Q_{2m}|+\hat{\mathfrak{s}}(2m))}{2mb\hat{\gamma}_m^\beta} \right) + \hat{\mathfrak{s}}(2m)}{m}}$$

$$+ c_1 \bar{\ell} \frac{\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}6(|Q_{2m}|+\hat{\mathfrak{s}}(2m))}{2mb\hat{\gamma}_m^\beta} \right) + \hat{\mathfrak{s}}(2m)}{m}$$

$$\leq 8c_1 \sqrt{b\hat{\gamma}_m^\beta \frac{\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}(|Q_{2m}|+\hat{\mathfrak{s}}(2m))}{2mb\hat{\gamma}_m^\beta} \right) + \hat{\mathfrak{s}}(2m)}{2m}}$$

$$+ 8c_1 \bar{\ell} \frac{\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}\left(\frac{\bar{\ell}(|Q_{2m}|+\hat{\mathfrak{s}}(2m))}{2mb\hat{\gamma}_m^\beta} \right) + \hat{\mathfrak{s}}(2m)}{2m}.$$

Thus, if we take $c_0 = 8c_1$ in the definition of $\hat{T}_\ell$ in (12.34), then we have

$$\tilde{U}_\ell \left(V^{(2m)}_{D_{2m}}; \mathcal{P}_{XY}, m, \hat{\mathfrak{s}}(2m) \right) \wedge \bar{\ell} \leq \frac{|Q_{2m}| \vee 1}{m} \left(\hat{T}_\ell \left(V^{(2m)}; Q_{2m}, 2m \right) \wedge \bar{\ell} \right).$$

276

Furthermore, (12.63) implies $|Q_{2m}| \leq \hat{\mathfrak{s}}(2m)+2em\mathcal{P}(D_{2m})$. In particular, if $\hat{\mathfrak{s}}(2m) > 2em\mathcal{P}(D_{2m})$, then

$$\frac{|Q_{2m}| \vee 1}{m}\left(\hat{T}_\ell\left(V^{(2m)}; Q_{2m}, 2m\right) \wedge \bar{\ell}\right) \leq \frac{\hat{\mathfrak{s}}(2m) + 2em\mathcal{P}(D_{2m})}{m}\bar{\ell} \leq \frac{2\hat{\mathfrak{s}}(2m)\bar{\ell}}{m},$$

and taking any $c_2 \geq 4$ guarantees this last quantity is at most $\tilde{T}_\ell(2m)$. On the other hand, if $\hat{\mathfrak{s}}(2m) \leq 2em\mathcal{P}(D_{2m})$, then $|Q_{2m}| \leq 4em\mathcal{P}(D_{2m})$, and we have already established that $V^{(2m)} \subseteq \mathcal{F}(\mathcal{E}_\ell(\hat{\gamma}_m); \text{01})$, so that

$$\frac{|Q_{2m}| \vee 1}{m}\left(\hat{T}_\ell\left(V^{(2m)}; Q_{2m}, 2m\right) \wedge \bar{\ell}\right)$$

$$\leq 8c_1\sqrt{b\hat{\gamma}_m^\beta \frac{\text{vc}(\mathcal{G}_\mathcal{F})\text{Log}\left(\frac{\bar{\ell}3e\mathcal{P}(\text{DIS}(\mathcal{F}(\mathcal{E}_\ell(\hat{\gamma}_m); \text{01})))}{b\hat{\gamma}_m^\beta}\right) + \hat{\mathfrak{s}}(2m)}{2m}}$$

$$+ 8c_1\bar{\ell}\frac{\text{vc}(\mathcal{G}_\mathcal{F})\text{Log}\left(\frac{\bar{\ell}3e\mathcal{P}(\text{DIS}(\mathcal{F}(\mathcal{E}_\ell(\hat{\gamma}_m); \text{01})))}{b\hat{\gamma}_m^\beta}\right) + \hat{\mathfrak{s}}(2m)}{2m}. \quad (12.67)$$

If $\hat{\gamma}_m \geq \Psi_\ell(\varepsilon)$, then this is at most

$$8c_1\left(\sqrt{b\hat{\gamma}_m^\beta \frac{\text{vc}(\mathcal{G}_\mathcal{F})\text{Log}\left(3e\chi_\ell\bar{\ell}\right) + \hat{\mathfrak{s}}(2m)}{2m}} + \bar{\ell}\frac{\text{vc}(\mathcal{G}_\mathcal{F})\text{Log}\left(3e\chi_\ell\bar{\ell}\right) + \hat{\mathfrak{s}}(2m)}{2m}\right)$$

$$\leq 48c_1\left(\sqrt{b\hat{\gamma}_m^\beta \frac{\text{vc}(\mathcal{G}_\mathcal{F})\text{Log}\left(\chi_\ell\bar{\ell}\right) + \hat{\mathfrak{s}}(2m)}{2m}} + \bar{\ell}\frac{\text{vc}(\mathcal{G}_\mathcal{F})\text{Log}\left(\chi_\ell\bar{\ell}\right) + \hat{\mathfrak{s}}(2m)}{2m}\right).$$

For brevity, let $K = \frac{\text{vc}(\mathcal{G}_\mathcal{F})\text{Log}(\chi_\ell\bar{\ell})+\hat{\mathfrak{s}}(2m)}{2m}$. As argued above, $\hat{\gamma}_m \leq 2\tilde{T}_\ell(m)$, so that the right hand side of the above inequality is at most

$$48\sqrt{2}c_1\left(\sqrt{b\tilde{T}_\ell(m)^\beta K} + \bar{\ell}K\right).$$

Then since $\hat{\mathfrak{s}}(m) \leq 2\hat{\mathfrak{s}}(2m)$, the above expression is at most

$$48 \cdot 4c_1\sqrt{c_2}\left(\sqrt{b\left((bK)^{\frac{1}{2-\beta}} \vee \bar{\ell}K\right)^\beta K} + \bar{\ell}K\right). \quad (12.68)$$

If $\bar{\ell}K \leq (bK)^{\frac{1}{2-\beta}}$, then (12.68) is equal

$$48 \cdot 4c_1\sqrt{c_2}\left((bK)^{\frac{1}{2-\beta}} + \bar{\ell}K\right).$$

On the other hand, if $\bar{\ell}K > (bK)^{\frac{1}{2-\beta}}$, then (12.68) is equal

$$48 \cdot 4c_1\sqrt{c_2}\left(\sqrt{bK(\bar{\ell}K)^\beta} + \bar{\ell}K\right)$$

$$< 48 \cdot 4c_1\sqrt{c_2}\left(\sqrt{(\bar{\ell}K)^{2-\beta}(\bar{\ell}K)^\beta} + \bar{\ell}K\right) = 48 \cdot 8c_1\sqrt{c_2}\bar{\ell}K.$$

In all of the above cases, taking $c_2 = 9 \cdot 2^{14}c_1^2$ in the definition of $\tilde{T}_\ell$ yields

$$\frac{|Q_{2m}| \vee 1}{m}\left(\hat{T}_\ell\left(V^{(2m)}; Q_{2m}, 2m\right) \wedge \bar{\ell}\right) \leq \tilde{T}_\ell(2m).$$

This completes the inductive step, so that we have proven that the claim holds for all m' with $\log_2(m') \in \{1, \ldots, i_\varepsilon\}$.

Let $j_\ell = -\lceil\log_2(\bar{\ell})\rceil$, $\tilde{j}_\varepsilon = \lceil\log_2(1/\Psi_\ell(\varepsilon))\rceil$, and for each $j \in \{j_\ell, \ldots, \tilde{j}_\varepsilon\}$, let $s_j = \log_2\left(\frac{144(2+\tilde{j}_\varepsilon-j)^2}{\delta}\right)$, define

$$m'_j = 32c_2^2\left(b2^{j(2-\beta)} + \bar{\ell}2^j\right)\left(\mathrm{vc}(\mathcal{G}_\mathcal{F})\mathrm{Log}(\chi_\ell\bar{\ell}) + s_j\right),$$

and let $m_j = 2^{\lceil\log_2(m'_j)\rceil}$. Also define $m_{j_\ell-1} = 1$. Using this notation, we can now define the relevant values of the $\hat{\mathfrak{s}}$ function as follows. For each $j \in \{j_\ell, \ldots, \tilde{j}_\varepsilon\}$, and each $m \in \{m_{j-1}+1, \ldots, m_j\}$ with $\log_2(m) \in \mathbb{N}$, define

$$\hat{\mathfrak{s}}(m) = \log_2\left(\frac{16\log_2(4m_j/m)^2(2+\tilde{j}_\varepsilon-j)^2}{\delta}\right).$$

In particular, taking $i_\varepsilon = \log_2(m_{\tilde{j}_\varepsilon})$, we have that $2\tilde{T}_\ell(2^{i_\varepsilon-1}) \leq \Psi_\ell(\varepsilon)$, so that on the event $\bigcap_{i=1}^{i_\varepsilon-1} E_{2^i} \cap E''_{2^i+1}$, if we have $2^{i_\varepsilon} \in S$, then $\hat{h} \in V^{(2^{i_\varepsilon})} \subseteq \mathcal{F}(\mathcal{E}_\ell(2\tilde{T}_\ell(2^{i_\varepsilon-1}) \vee \Psi_\ell(\varepsilon)); 01) = \mathcal{F}(\mathcal{E}_\ell(\Psi_\ell(\varepsilon)); 01) \subseteq \mathcal{F}(\Psi_\ell^{-1}(\Psi_\ell(\varepsilon)); 01) = \mathcal{F}(\varepsilon; 01)$, so that $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$.

Furthermore, we established above that, on the event $\bigcap_{i=1}^{i_\varepsilon-1} E_{2^i} \cap E''_{2^i+1}$, for every $j \in \{j_\ell, \ldots, \tilde{j}_\varepsilon\}$ with $m_j \in S$, and every $m \in \{m_{j-1}+1, \ldots, m_j\}$ with $\log_2(m) \in \mathbb{N}$, $V^{(m)} \subseteq \mathcal{F}(\mathcal{E}_\ell(2\tilde{T}_\ell(m/2) \vee \Psi_\ell(\varepsilon)); 01) \subseteq \mathcal{F}(\mathcal{E}_\ell(2\tilde{T}_\ell(m_{j-1}) \vee \Psi_\ell(\varepsilon)); 01)$. Noting that $2\tilde{T}_\ell(m_{j-1}) \leq 2^{1-j}$, we have

$$\sum_{m \in S: m \leq m_{\tilde{j}_\varepsilon}} |Q_m| \leq \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \sum_{m=m_{j-1}+1}^{m_j} \mathbb{I}_{\mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{1-j}); 01))}(X_m).$$

A Chernoff bound implies that, on an event E' of probability at least $1 - \delta/2$, the right hand side of the above inequality is at most

$$\log_2(2/\delta) + 2e \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} (m_j - m_{j-1}) \mathcal{P}(\mathrm{DIS}(\mathcal{F}(\mathcal{E}_\ell(2^{1-j}); {}_{01})))$$

$$\leq \log_2(2/\delta) + 2e \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} m_j \mathcal{P}(\mathrm{DIS}(\mathcal{F}(\Psi_\ell^{-1}(2^{1-j}); {}_{01}))).$$

By essentially the same reasoning used in the proof of Theorem 12.16, the right hand side of this inequality is

$$\lesssim a\theta\varepsilon^\alpha \left(\frac{b(A_1 + \mathrm{Log}(B_1))B_1}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}(A_1 + \mathrm{Log}(C_1))C_1}{\Psi_\ell(\varepsilon)} \right).$$

Since

$$m_{\tilde{j}_\varepsilon} \lesssim \left(\frac{b}{\Psi_\ell(\varepsilon)^{2-\beta}} + \frac{\bar{\ell}}{\Psi_\ell(\varepsilon)} \right) A_1,$$

the conditions on u and n stated in Theorem 12.16 (with an appropriate constant c) suffice to guarantee $\mathrm{er}(\hat{h}) - \mathrm{er}(h^*) \leq \varepsilon$ on the event $E' \cap \bigcap_{i=1}^{i_\varepsilon - 1} E_{2^i} \cap E''_{2^{i+1}}$. Finally, the proof is completed by noting that a union bound implies the event $E' \cap \bigcap_{i=1}^{i_\varepsilon - 1} E_{2^i} \cap E''_{2^{i+1}}$ has probability at least

$$1 - \frac{\delta}{2} - \sum_{i=1}^{i_\varepsilon - 1} 2^{1-\hat{\mathfrak{s}}(2^{i+1})} + 6e^{-\hat{\mathfrak{s}}(2^i)}$$

$$\geq 1 - \frac{\delta}{2} - \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \sum_{i=\log_2(m_{j-1})+1}^{\log_2(m_j)} \frac{\delta}{2(2 + \log_2(m_j) - i)^2(2 + \tilde{j}_\varepsilon - j)^2}$$

$$\geq 1 - \frac{\delta}{2} - \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \sum_{k=0}^{\infty} \frac{\delta}{2(2 + k)^2(2 + \tilde{j}_\varepsilon - j)^2}$$

$$\geq 1 - \frac{\delta}{2} - \sum_{j=j_\ell}^{\tilde{j}_\varepsilon} \frac{\delta}{2(2 + \tilde{j}_\varepsilon - j)^2} \geq 1 - \frac{\delta}{2} - \sum_{t=0}^{\infty} \frac{\delta}{2(2 + t)^2} \geq 1 - \delta.$$

Note that, as in Theorem 12.16, the function $\hat{\mathfrak{s}}$ in this proof has a direct dependence on a, α, and χ_ℓ, in addition to b and β. As before, with an alternative definition of $\hat{\mathfrak{s}}$, similar to that mentioned in the discussion following Theorem 12.16, it is possible to remove this dependence, at the expense of the same logarithmic factors mentioned above.

Chapter 13

Online Allocation and Pricing with Economies of Scale

Abstract

[1]Allocating multiple goods to customers in a way that maximizes some desired objective is a fundamental part of Algorithmic Mechanism Design. We consider here the problem of offline and online allocation of goods that have economies of scale, or decreasing marginal cost per item for the seller. In particular, we analyze the case where customers have unit-demand and arrive one at a time with valuations on items, sampled iid from some unknown underlying distribution over valuations. Our strategy operates by using an initial sample to learn enough about the distribution to determine how best to allocate to future customers, together with an analysis of structural properties of optimal solutions that allow for uniform convergence analysis. We show, for instance, if customers have binary valuations over items, and the goal of the allocator is to give each customer an item he or she values, we can efficiently produce such an allocation with cost at most a constant factor greater than the minimum over such allocations in hindsight, so long as the marginal costs do not decrease too rapidly. We also give a bicriteria approximation

---

[1]This chapter is based on joint work with Avrim Blum and Yishay Mansour.

to social welfare for the case of more general valuation functions when the allocator is budget constrained.

13.1 Introduction

Imagine it is the Christmas season, and Santa Claus is tasked with allocating toys. There is a sequence of children coming up with their Christmas lists of toys they want. Santa wants to give each child some toy from his or her list (for simplicity, assume all children have been good this year). But of course, even Santa Claus has to be cost-conscious, so he wants to perform this allocation of toys to children at a near-minimum cost to himself (call this the Thrifty Santa Claus Problem). Now if it was the case that every toy had a fixed price, this would be easy: simply allocate to each child the cheapest toy on his or her list and move on to the next child. But here we are interested in the case where goods have economies of scale. For example, producing a millon toy cars might be cheaper than a million times the cost of producing one toy car. Thus, even if producing a single toy car is more expensive than a single Elmo doll, if a much larger number of children want the toy car than the Elmo doll, the minimum-cost allocation might give toy cars to many children, even if some of them also have the Elmo doll on their lists.

The problem faced by Santa (or by any allocator that must satisfy a collection of disjunctive constraints in the presence of economies of scale) makes sense in both offline and online settings. In the offline setting, in the extreme case of goods such as software where all the cost is in the first copy, this is simply weighted set-cover, admitting a $\Theta(\log n)$ approximation to the minimum-cost allocation. We will be interested in the online case where customers are iid samples from some arbitrary distribution over subsets of item-set $\mathcal{I}$ (i.e., Christmas lists), where the allocator must make allocation decisions online, and where the marginal cost of goods does not decrease so sharply. We show that for a range of cost curves, including the case that the marginal cost of copy t of an item is $t^{-\alpha}$, for some $\alpha \in [0, 1)$, we will be able to get a constant-factor approximation so long as the number of customers is sufficiently large compared to the number of items.

One basic observation we show is that, if the marginal costs are non-increasing, there is always an optimal allocation that can be described as an ordering of the possible toys, so that as each child comes, Santa simply gives the child the first toy in the ordering that appears on the child's list. Another observation we prove is that, if the marginal costs do not drop too quickly, then if we are given the lists of all the children before determining the allocation, we can efficiently find an allocation that is within a constant factor of the minimum-cost allocation, as opposed to the logarithmic factor required for the set-cover problem. Since, however, the problem we are interested in does not supply the lists before the allocations, but rather requires a decision for each child in sequence, we rely on the iid assumption and use ideas from machine learning, as follows: after processing a small initial number of children (with no nontrivial guarantees on allocation costs for these), we take their wish lists as representative of the future children, and find the optimal solution (in hindsight) for those, while treating each of these children as representing many future children (supposing we know the total number of children ahead of time). We then take the ordered list of toys from this solution, and allocate according to this preference ordering in the future (allocating to each child the earliest toy in the ordering that is also on his or her list). We show that, as long as we take a sufficiently large number of initial children, this procedure will find an ordering that will be near-optimal for allocating to the remaining children.

More generally, we can imagine the case where, rather than simple lists of items, the lists also provide valuations for each item, and we are interested in the trade-off between maximizing the total of valuations for allocated items while minimizing the total cost of the allocation. In this case, we might think of the allocator as being a large company with many different projects, and each project has some valuations over different resources (e.g., types of laptops for employees involved in that project), where it could use one or another resource but prefers some resources over others. One natural quantity to consider in this context is the social welfare: the difference between the happiness (total of valuations for the allocation) minus the total cost of the allocation. In this case, it turns out the optimal allocation rule can be described by a pricing scheme. In

another words, whatever the optimal allocation is, there always exist prices such that if the buyers purchase what they most want at those prices, they will actually produce that allocation. We note that, algorithmically, this is a harder problem than the list-based problem (which corresponds to binary valuations).

Aside from social welfare, it is also interesting to consider a variant in which we have a budget constraint, and are interested in maximizing the total valuation of the allocation, subject to that budget constraint on the total cost of the allocation. It turns out this latter problem can be reduced to a problem known as the weighted budget maximum coverage problem. Technically, this problem is originally formulated for the case in which the marginal cost of a given item drops to zero after the first item of that type is allocated (as in the set cover reduction mentioned above); however, viewed appropriately, we are able to formulate this reduction for arbitrary decreasing marginal cost functions. What we can then do is run an algorithm for the weighted budget maximum coverage problem, and then convert the solution into a pricing. As before, this strategy will be effective for the offline problem, in which all of the valuations are given ahead of time. However, we can extend it to the online setting with iid valuation functions by generating a pricing based on an appropriately-sized initial sample of valuation functions, and then apply that pricing to sequentially generate allocations for the remaining valuations. Again, as long as the marginal costs are not decreasing too rapidly, we can obtain an allocation strategy for which the sum of valuations of the allocated items will be within a constant factor of the maximum possible, subject to the budget constraint on the cost.

13.1.1   Our Results and Techniques

We consider this problem under two, related, natural objectives. In the first (the "thrifty Santa Claus" objective) we assume customers have binary $\{0, 1\}$ valuations, and the goal of the seller is to give each customer a toy of value 1, but in such a way that minimizes the total cost to the seller. We show that so long as the number of buyers n is large compared to the number of items r, and

so long as the marginal costs do not decrease too rapidly (e.g., a rate $1/t^\alpha$ for some $0 \le \alpha < 1$), we can efficiently perform this allocation task with cost at most a constant factor greater than that of the optimal allocation of items in hindsight. Note that if costs decrease much more rapidly, then even if all customers' valuations were known up front, we would be faced with (roughly) a set-cover problem and so one could not hope to achieve cost $o(\log n)$ times optimal. The second objective we consider, which we apply to customers of arbitrary unit-demand valuation, is that of maximizing total social welfare of customers subject to a cost bound on the seller; for this, we also give a strategy that is constant-competitive with respect to the optimal allocation in hindsight.

Our algorithms operate by using initial buyers to learn enough about the distribution to determine how best to allocate to the future buyers. In fact, there are two main technical parts of our work: the sample complexity and the algorithmic aspects. From the perspective of sample complexity, one key component of this analysis is examining how complicated the allocation rule needs to be in order to achieve good performance, because simpler allocation rules require fewer samples in order to learn. We do this by providing a characterization of what the optimal strategies look like. For example, for the thrifty Santa Claus version, we show that the optimal solution can be assumed wlog to have a simple permutation structure. In particular, so long as the marginal costs are nonincreasing, there is always an optimal strategy in hindsight of this form: order the items according to some permutation and for each bidder, give it the earliest item of its desire in the permutation. This characterization is used inside both our sample complexity results and our algorithmic guarantees. Specifically, we prove that for cost function $\text{cost}(t) = \sum_{\tau=1}^{t} 1/\tau^\alpha$, for $\alpha \in [0, 1)$, running greedy weighted set cover incurs total cost at most $\frac{1}{1-\alpha}\text{OPT}$. More generally, if the average cost is within some factor of the marginal cost, we have a greedy algorithm that achieves constant approximation ratio. To allocate to new buyers, we simply give it the earliest item of its desire in the learnt permutation. For the case of general valuations, we give a characterization showing that the optimal allocation rule in terms

of social welfare can be described by a pricing scheme. That is, there exists a pricing scheme such that if buyers purchased their preferred item at these prices, the optimal allocation would result. Algorithmically, we show that we can reduce to a weighted budgeted maximum coverage problem with single-parameter demand for which there is a known constant-approximation-ratio algorithm [Khuller, Moss, and Naor, 1999].

13.1.2 Related Work

In this work we focus on the case of decreasing marginal cost. There have been a large body of research devoted to unlimited supply, which is implicitly constant marginal cost (e.g., [Nisan, Roughgarden, Tardos, and Vazirani, 2007] Chapter 13), where the goal is to achieve a constant competitive ratio in both offline and online models. The case of increasing marginal cost was studies in [Blum, Gupta, Mansour, and Sharma, 2011] where constant competitive ratio where given.

We analyze an online setting where buyers arrive one at a time, sampled iid from some unknown underlying distribution over valuations. Other related online problems with stochastic inputs such as matching problems have been studied in ad auctions [Goel and Mehta, 2008, Mehta, Saberi, Vazirani, and Vazirani, 2007]. Algorithmically, our work is related to the online set cover body of work where [Alon, Awerbuchy, Azarz, Buchbinder, and Naor, 2009] gave the first $O(\log m \log n)$ competitive algorithm (here n is the number of elements in the ground set and m is size of a family of subsets of the ground set). The problems we study are also related to online matching problems [Devanur and Hayes, 2009, Devanur and Jain, 2012, Karp, Vazirani, and Vazirani, 1990] in the iid setting; however our problem is a bit like the "opposite" of online matching in that the cumulative cost curve for us is concave rather than convex.

13.2 Model, Definitions, and Notation

We have a set $\mathcal{I}$ of r items. We have a set $N = \{1, \ldots, n\}$ indexing n unit demand buyers. Our setting can then generally be formalized in the following terms.

13.2.1 Utility Functions

Each buyer $j \in N$ has a weight $u_{j,i}$ for each item $i \in \mathcal{I}$. We suppose the vectors $u_{j,\cdot}$ are sampled i.i.d. according to a fixed (but arbitrary and unknown) distribution. In the *online* setting we are interested in, the buyers' weight vectors $u_{j,\cdot}$ are observed in sequence, and for each one (before observing the next) we are required to allocate a set of items $T_j \subseteq \mathcal{I}$ to that buyer. The *utility* of buyer j for this allocation is then defined as $u_j(T_j) = \max_{i \in T_j} u_{j,i}$. A few of our results consider a slight variant of this model, in which we are only required to begin allocating goods after some initial $o(n)$ number of customers has been observed (to whom we may allocate items retroactively).

This general setting is referred to as the *weighted unit demand* setting. We will also be interested in certain special cases of this problem. In particular, many of our results are for the *uniform unit demand* setting, in which every $j \in N$ and $i \in \mathcal{I}$ have $u_{j,i} \in \{0, 1\}$. In this case, we may refer to the set $S_j = \{i \in \mathcal{I} : u_{j,i} = 1\}$ as the list of items buyer j *wants* (one of).

13.2.2 Production cost

We suppose there are *cumulative cost functions* $\mathrm{cost}_i : \mathbb{N} \to [0, \infty]$ for each item $i \in \mathcal{I}$, where for $t \in \mathbb{N}$, the value of $\mathrm{cost}_i(t)$ represents the cost of producing t copies of item i. We suppose each $\mathrm{cost}_i(\cdot)$ is nondecreasing.

We would like to consider the case of *decreasing marginal cost*, where $t \mapsto \mathrm{cost}_i(t + 1) - \mathrm{cost}_i(t)$ is nonincreasing for each $i \in \mathcal{I}$.

A natural class of decreasing marginal costs we will be especially interested in are of the

form $t^{-\alpha}$ for $\alpha \in [0, 1)$. That is, $\mathrm{cost}_i(t) = c_0 \sum_{\tau=1}^{t} \tau^{-\alpha}$.

13.2.3 Allocation problems

After processing the n buyers, we will have allocated some set of items T, consisting of $m_i(T) = \sum_{j \in N} \mathbb{I}_{T_j}(i)$ copies of each item $i \in \mathcal{I}$. We are then interested in two quantities in this setting: the *total (production) cost* $\mathrm{cost}(T) = \sum_{i \in \mathcal{I}} \mathrm{cost}_i(m_i(T))$ and the *social welfare* $SW(T) = \sum_{j \in N} u_j(T_j)$.

We are interested in several different objectives within this setting, each of which is some variant representing the trade-off between reducing total production cost while increasing social welfare.

In the *allocate all* problem, we have to allocate to each buyer $j \in N$ one item $i \in S_j$ (in the uniform demand setting): that is, $SW(T) = n$. The goal is to minimize the total cost $\mathrm{cost}(T)$, subject to this constraint.

The *allocate with budget* problem requires our total cost to never exceed a given limit b (i.e., $\mathrm{cost}(T) \leq b$). Subject to this constraint, our objective is to maximize the social welfare $SW(T)$. For instance, in the uniform demand setting, this corresponds to maximizing the number of satisfied buyers (that get an item from their set S_j).

The objective in the *maximize social surplus* problem is to maximize the difference of the social welfare and the total cost (i.e., $SW(T) - \mathrm{cost}(T)$).

13.3 Structural Results and Allocation Policies

We now present several results about the structure of optimal (and non-optimal but "reasonable") solutions to allocation problems in the setting of decreasing marginal costs. These will be important in our sample-complexity analysis because they allow us to focus on allocation policies that have inherent complexity that depends only on the number of *items* and not on the number of

customers, allowing for the use of uniform convergence bounds. That is, a small random sample of customers will be sufficient to uniformly estimate the performance of these policies over the full set of customers.

13.3.1 Permutation and pricing policies

A *permutation policy* has a permutation π over $\mathcal{I}$ and is applicable in the case of uniform unit demand. Given buyer j arriving, we allocate to him the minimal (first) demanded item in the permutation, i.e., $\arg\min_{i \in S_j} \pi(i)$. A *pricing policy* assigns a price price_i to each item i and is applicable to general quasilinear utility functions. Given buyer j arriving, we allocate to him whatever he wishes to purchase at those prices, i.e., $\arg\max_{T_j} u_j(T_j) - \sum_{i \in T_j} \text{price}_i$.[^2]

We will see below that for uniform unit demand buyers, there always exists a permutation policy that is optimal for the allocate-all task, and for general quasilinear utilities there always exists a pricing policy that is optimal for the task of maximizing social surplus. We will also see that for weighted unit demand buyers, there always exists a pricing policy that is optimal for the allocate-with-budget task; moreover, for any even non-optimal solution (e.g., that might be produced by a polynomial-time algorithm) there exists a pricing policy that sells the same number of copies each item and has social welfare at least as high (and can be computed in polynomial time given the initial solution).

13.3.2 Structural results

Theorem 13.1. *For general quasilinear utilities, any allocation that maximizes social surplus can be produced by a pricing policy. That is, if $\mathcal{T} = \{T_1, \ldots, T_n\}$ is an allocation maximizing $SW(T) - \text{cost}(T)$ then there exist prices $\text{price}_1, \ldots, \text{price}_r$ such that buyers purchasing their most-demanded bundle recovers $\mathcal{T}$, assuming that the marginal cost function is strictly decreas-*

[^2]: When more that one subset is applicable, we assume we have the freedom to select any such set.

ing.[3]

Proof. Consider the optimal allocation OPT. Define price_i to be the marginal cost of the next copy of item i under OPT, i.e., $\mathrm{price}_i = \mathrm{cost}_i(\#_i(\mathrm{OPT})+1)$. Suppose some buyer j is assigned set T_j in OPT but prefers set T_j' under these prices. Then,

$$u_j(T_j') - \sum_{i\in T_j'} \mathrm{price}_i \geq u_j(T_j) - \sum_{i\in T_j} \mathrm{price}_i,$$

which implies

$$u_j(T_j') - u_j(T_j) + \sum_{i\in T_j\setminus T_j'} \mathrm{price}_i - \sum_{i\in T_j'\setminus T_j} \mathrm{price}_i \geq 0. \tag{13.1}$$

Now, consider modifying OPT by replacing T_j with T_j'. This increases buyer j's utility by $u_j(T_j') - u_j(T_j)$, incurs an extra purchase cost *exactly* $\sum_{i\in T_j'\setminus T_j} \mathrm{price}_i$ and a savings of strictly more than $\sum_{i\in T_j\setminus T_j'} \mathrm{price}_i$ (because marginal costs are decreasing). Thus, by (13.1) this would be a strictly preferable allocation, contradicting the optimality of OPT. $\qquad\square$

Corollary 13.2. *For uniform unit demand buyers there exists an optimal allocation that is a permutation policy, for the* allocate all *task.*

Proof. Imagine each buyer j had valuation v_{max} on items in S_j where v_{max} is greater than the maximum cost of any single item. The allocation OPT that maximizes social surplus would then minimize cost subject to allocating exactly one item to each buyer and therefore would be optimal for the allocate-all task. Consider the pricing associated to this allocation given by Theorem 13.1. Since each buyer j is uniform unit demand, he will simply purchase the cheapest item in S_j. Therefore, the permutation π that orders items according to increasing price according to the prices of Theorem 13.1 will produce the same allocation. $\qquad\square$

We now present a structural statement that will be useful for the allocate-with-budget task.

---

[3]If the marginal cost function is only non-increasing, we can have the same result, assuming we can select between the utility maximizing bundles.

Theorem 13.3. *For weighted unit-demand buyers, for any allocation $\mathcal{T}$ there exists a pricing policy that allocates the same* multiset *of items T (or a subset of T) and has social welfare at least as large as $\mathcal{T}$. Moreover, this pricing can be computed efficiently from $\mathcal{T}$ and the buyers' valuations.*

Proof. Let T be the multiset of items allocated by $\mathcal{T}$. Weighted unit-demand valuations satisfy the gross-substitutes property, so by the Second Welfare Theorem (e.g., see [Nisan, Roughgarden, Tardos, and Vazirani, 2007] Theorem 11.15) there exists a Walrasian equilibrium: a set of prices for the items in T that clears the market. Moreover, these prices can be computed efficiently from demand queries (e.g., [Nisan, Roughgarden, Tardos, and Vazirani, 2007], Theorem 11.24), which can be evaluated efficiently for weighted unit-demand buyers. Furthermore, these prices must assign all copies of the *same* item in T the same price (else the pricing would not be an equilibrium) so it corresponds to a legal pricing policy. Thus, we have a legal pricing such that if all buyers were shown only the items represented in T, at these prices, then the market would clear perfectly (breaking any ties in our favor). We can address the fact that there may be items not represented in T (i.e., they had zero copies sold) by simply setting their price to infinity. Finally, by the First Welfare Theorem (e.g., [Nisan, Roughgarden, Tardos, and Vazirani, 2007] Theorem 11.13), this pricing maximizes social welfare over all allocations of T, and therefore achieves social welfare at least as large as $\mathcal{T}$, as desired. □

The above structural results will allow us to use the following sketch of an online algorithm. First sample an initial set of ℓ buyers. Then, for the allocate-all problem, compute the best (or approximately best) permutation policy according to the empirical frequencies given by the sample. Or, for the allocate-with budget task, compute the best (or approximately best) allocation according to these empirical frequencies and convert it into a pricing policy. Then run this permutation or pricing policy on the remainder of the customers. Finally, using the fact that these policies have low complexity (they are lists or vectors in a space that depends only on the number of items and not on the number of buyers) compute the size of initial sample needed to

ensure that the estimated performance is close to true performance uniformly over all policies in the class.

13.4 Uniform Unit Demand and the Allocate-All problem

Here we consider the allocate-all problem for the setting of uniform unit demand. For intuition, we begin by considering the following simple class of decreasing marginal cost curves.

Definition 13.4. *We say the cost function* $\mathrm{cost}(t)$ *is* α*-poly if the marginal cost of item* t *is* $1/t^\alpha$ *for* $\alpha \in [0, 1)$. *That is,* $\mathrm{cost}(t) = \sum_{\tau=1}^{t} 1/\tau^\alpha$.

Theorem 13.5. *If each cost function is* α*-poly, then there exists an efficient offline algorithm that given a set* X *of buyers produces a permutation policy that incurs total cost at most* $\frac{1}{1-\alpha}\mathrm{OPT}$.

Proof. We run the greedy set-cover algorithm. Specifically, we choose the item desired by the most buyers and put it at the top of the permutation π. We then choose the item desired by the most buyers who did not receive the first item and put it next, and so on. For notational convenience assume π is the identity, and let $\mathcal{S}_i$ denote the set of buyers that receive item i, For any set $\mathcal{S} \subseteq X$, let $\mathrm{OPT}(\mathcal{S})$ denote the cost of the optimal solution to the subproblem $\mathcal{S}$ (i.e., the problem in which we are only required to cover buyers in $\mathcal{S}$). Clearly $\mathrm{OPT}(\mathcal{S}_r) = \mathrm{cost}(|\mathcal{S}_r|) = \sum_{\tau=1}^{|\mathcal{S}_r|} 1/\tau^\alpha \geq \sum_{t=1}^{|\mathcal{S}_r|} \int_1^{|\mathcal{S}_t|} x^{-\alpha} \mathrm{d}x = \frac{1}{1-\alpha}|\mathcal{S}_r|^{1-\alpha} - 1$, since any solution using more than one set to cover the elements of $\mathcal{S}_r$ has at least as large a cost.

Now, for the purpose of induction, suppose that some $k \in \{2, \ldots, r\}$ has $\mathrm{OPT}(\bigcup_{t=k}^{r} \mathcal{S}_t) \geq \sum_{t=k}^{r} |\mathcal{S}_t|^{1-\alpha}$. Then, since $\mathcal{S}_{k-1}$ was chosen to be the largest subset of $\bigcup_{t=k-1}^{r} \mathcal{S}_t$ that can be covered by a single item, it must be that the sets used by any allocation for the $\bigcup_{t=k-1}^{r} \mathcal{S}_t$ subproblem achieving $\mathrm{OPT}(\bigcup_{t=k-1}^{r} \mathcal{S}_t)$ have size at most $|\mathcal{S}_{k-1}|$, and thus the marginal costs for each of the elements of $\mathcal{S}_{k-1}$ in the $\mathrm{OPT}(\bigcup_{t=k-1}^{r} \mathcal{S}_t)$ solution is at least $1/|\mathcal{S}_{k-1}|^\alpha$.

This implies $\mathrm{OPT}(\bigcup_{t=k-1}^{r} \mathcal{S}_t) \geq \mathrm{OPT}(\bigcup_{t=k}^{r} \mathcal{S}_t) + \sum_{x \in \mathcal{S}_{k-1}} 1/|\mathcal{S}_{k-1}|^\alpha = \mathrm{OPT}(\bigcup_{t=k}^{r} \mathcal{S}_t) + |\mathcal{S}_{k-1}|^{1-\alpha}$. By the inductive hypothesis, this latter expression is at least as large as $\sum_{t=k-1}^{r} |\mathcal{S}_t|^{1-\alpha}$.

By induction, this implies $\text{OPT}(X) = \text{OPT}(\bigcup_{t=1}^{r} \mathcal{S}_t) \geq \sum_{t=1}^{r} |\mathcal{S}_t|^{1-\alpha}$. On the other hand, the total cost incurred by the greedy algorithm is $\sum_{t=1}^{r} \sum_{\tau=1}^{|\mathcal{S}_r|} 1/\tau^\alpha \leq \sum_{t=1}^{r} \int_{0}^{|\mathcal{S}_t|} x^{-\alpha} \mathrm{d}x = \frac{1}{1-\alpha} \sum_{t=1}^{r} |\mathcal{S}_t|^{1-\alpha}$. By the above argument, this is at most $\frac{1}{1-\alpha} \text{OPT}(X)$. $\qquad\square$

More general cost curves We can generalize the above result to a natural class of smoothly decreasing cost curves. Define the average cost of item i given to set $\mathcal{S}_i$ of buyers as $AvgC(i, |\mathcal{S}_i|) = \frac{\text{cost}(|\mathcal{S}_i|)}{|\mathcal{S}_i|}$. Define the marginal cost $MarC(i, t) = \text{cost}_i(t) - \text{cost}_i(t-1)$. Here is a greedy algorithm.

Algorithm: $GreedyGeneralCost(\mathcal{S})$

0. $\quad i = \arg\min AvgC(i, |\mathcal{S}_i|)$

1. $\quad$ Call $GreedyGeneralCost(\mathcal{S} - \mathcal{S}_i)$

We make the following assumption:

Assumption 13.6. $\forall i, t, \; AvgC(i, t) \leq \beta MarC(i, t)$, *for some* $\beta > 0$.

For example, for the case of an α-poly cost, we have: $MarC(t) = \frac{1}{t^\alpha}$ and $AvgC = \frac{1}{t} \sum_{\tau=1}^{t} \frac{1}{\tau^\alpha} \approx \frac{t^{-\alpha}}{1-\alpha}$; so, therefore we have $\beta = \frac{1}{1-\alpha}$.

Theorem 13.7. *The algorithm GreedyGeneralCost achieves approximation ratio β.*

Proof. Order the elements in the order that GreedyGeneralCost allocates them. Let N_j be the set of consumers that receive item j, and $N = \cup N_j$ in GreedyGeneralCost. For consumer i let $item_{opt}(i)$ be the item that OPT allocates to consumer i. Let $\ell_{opt}(j)$ be the number of consumers that are allocated item j. By Assumption 13.6 we have $MarC(j, l) \leq AvgC(j, l) \leq \beta MarC(j, l)$ (the first inequality is due to having decreasing marginal cost).

We would like to consider the influence of the consumers in N_1 on the cost of OPT. Let

$$
\begin{aligned}
\text{OPT}(N) - \text{OPT}(N - N_1) &\geq \sum_{i \in N_1} MarC(item_{opt}(i), \ell_{opt}(item_{opt}(i))) \\
&\geq \sum_{i \in N_1} \tfrac{1}{\beta} AvgC(item_{opt}(i), \ell_{opt}(item_{opt}(i))) \\
&\geq \tfrac{1}{\beta}|N_1| AvgC(1, |N_1|) = \tfrac{1}{\beta} GreedyCost(N_1)
\end{aligned}
$$

The first inequality follows since taking the final marginal cost can only reduce the cost (decreasing marginal cost). The second inequality follows from Assumption 13.6. The third inequality follows since GreedyGeneralCost selects the lowest average cost of any allocated item .

We can now continue inductively. Let $T_0 = N$, $T_1 = N - N_1$, and $T_i = T_{i-1} - N_i$. We can show similarly that,

$$
\text{OPT}(T_{i-1}) - \text{OPT}(T_i) \geq \frac{1}{\beta} GreedyCost(N_i)
$$

Summing over all i we have

$$
\begin{aligned}
OPT(T) - OPT(\emptyset) = \sum_i \text{OPT}(T_{i-1}) - \text{OPT}(T_i) &\geq \frac{1}{\beta} \sum_i GreedyCost(N_i) \\
&= \frac{1}{\beta} GreedyCost(N)
\end{aligned}
$$

$\square$

Corollary 13.8. *If the cost function is α-poly, then for $\beta = \frac{1}{1-\alpha}$, Assumption 13.6 holds. Thus* $\frac{\text{GreedyCost}(S_j)}{\text{OPTCost}(S_j)} \leq \frac{1}{1-\alpha}$.

Additionally, the following property is satisfied for these β-nice cost functions.

Lemma 13.9. *For* cost *satisfying Assumption 13.6, $\forall x \in \mathbb{N}$, $\forall \epsilon \in (0,1)$, $\forall i \leq r$, $\text{cost}_i(\epsilon x) \leq \epsilon^{\log_2(1+\frac{1}{2\beta})} \text{cost}_i(x)$.*

Proof. By the fact that marginal costs are non-negative, $AvgC(2\epsilon x) \geq \text{cost}_i(\epsilon x)/(2\epsilon x)$. Therefore, by Assumption 13.6, $MarC(2\epsilon x) \geq \text{cost}_i(\epsilon x)/(2\epsilon x\beta)$. By the decreasing marginal cost property, we have

$$
\text{cost}_i(2\epsilon x) \geq \text{cost}_i(\epsilon x) + \epsilon x MarC(2\epsilon x) \geq \text{cost}_i(\epsilon x) + \text{cost}_i(\epsilon x)/(2\beta) = (1 + \frac{1}{2\beta})\text{cost}_i(\epsilon x).
$$

Applying this argument $\log_2(1/\epsilon)$ times, we have

$$\mathrm{cost}_i(x) \geq (1 + \frac{1}{2\beta})^{\log_2(1/\epsilon)} \mathrm{cost}_i(\epsilon x) = (\frac{1}{\epsilon})^{\log_2(1+\frac{1}{2\beta})} \mathrm{cost}_i(\epsilon x).$$

Multiplying both sides by $\epsilon^{\log_2(1+\frac{1}{2\beta})}$ completes the proof. $\qquad\square$

13.4.1 Generalization Result

Say n is the total number of customers; ℓ is the size of subsample where we do estimate on; r is the total number of items; $\alpha \in (0,1]$ is some constant, and the cost is α-poly, so that $\mathrm{cost}(t) = \sum_{\tau=1}^t 1/\tau^\alpha \simeq \int_0^t y^{-\alpha} \mathrm{d}y = \left[\frac{y^{1-\alpha}}{1-\alpha}\right]_0^t = \frac{t^{1-\alpha}}{1-\alpha}$. We have the following generalization result:

Theorem 13.10. *Suppose $n \geq \ell$ and the cost function is α-poly. With probability at least $1-\delta^{(\ell)}$, for any permutations Π,*

$$\mathrm{cost}(\Pi, \ell)(1+\epsilon)^{-2} \left(\frac{n}{\ell}\right)^{1-\alpha} \leq \mathrm{cost}(\Pi, n) \leq \mathrm{cost}(\Pi, \ell)(1+\epsilon)^{2(1-\alpha)} \left(\frac{n}{\ell}\right)^{1-\alpha},$$

where $\delta^{(\ell)} = r2^r(\delta_1 + \delta_2 + \delta_3)$ and $\delta_1 = \exp\{-\epsilon^2 \left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}} n/3\}$, $\delta_2 = \exp\{-\epsilon^2\ell \left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}}/3\}$, $\delta_3 = \exp\{-\left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}} n\epsilon^2/2\}$.

Proof. Fix a permutation Π. Let π_j denote the event that a customer buys item Π_j and not covered by items Π_1 through Π_{j-1}. Namely, the probability that the consumer set of desired items include j and none of the items $1, \dots, j-1$. Let q_j denote $Pr[\pi_j]$, and let $\hat{q}_j$ denote the fraction of Π_j on the initial ℓ-sample.

Item j to is a "Low probability item" if $q_j < \left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}}$; and "High probability items" if $q_j \geq \left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}}$. Let the set "Low" include all "Low probability items"; and the set "High" include all "High probability items".

First we address the case of item j of low probability. The quantity of item j that we will sell is at most $\left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}} n(1+\epsilon)$ (Chernoff bound) with probability at least $1 - \delta_1$ with $\delta_1 = \exp\{-\epsilon^2 \left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}} n/3\}$. By a union bound, this holds for all low probability item j, with probability at least $1 - |\mathrm{Low}|\delta_1$.

Next, we suppose j has high probability. In this case, the quantity of item j we will sell is at most $q_j n(1 + \epsilon)$, with probability at least $1 - \exp\{-\epsilon^2 q_j n/3\} \geq 1 - \delta_1$. Again, a union bound implies this holds for all high probability j with probability at least $1 - |\text{High}|\delta_1$.

We have that (by Chernoff bounds), with probability at least $1 - \exp\{-\epsilon^2 \ell q_j/3\} \geq 1 - \delta_2$, we have $q_j/\hat{q}_j \leq (1 + \epsilon)$. A union bound implies this holds for all high probability j with probability $1 - r\delta_2$.

Furthermore, noting that $q_j n(1 + \epsilon) = \hat{q}_j n(1 + \epsilon)\frac{q_j}{\hat{q}_j}$, and upper bounding $\frac{q_j}{\hat{q}_j}$ by $1 + \epsilon$, we get that $q_j n(1 + \epsilon) \leq (1 + \epsilon)^2 \hat{q}_j n$, with probability $1 - \delta_2$. Thus,

$$
\begin{aligned}
\text{cost}(\Pi, n) &\leq \text{cost}(\text{Low}) + \text{cost}(\text{High}) \\
&\leq r\left(\left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}} n(1 + \epsilon)\right)^{1-\alpha} + \sum_{j \in \text{High}} \left((1 + \epsilon)^2 \hat{q}_j n\right)^{1-\alpha} \\
&\leq \epsilon(1 + \epsilon)^{1-\alpha} n^{1-\alpha} + (1 + \epsilon)^{2(1-\alpha)} n^{1-\alpha} \sum_{j \in \text{High}} (\hat{q}_j)^{1-\alpha}.
\end{aligned}
$$

Note that the total cost of all low probability items is at most ϵ-fraction of OPT which is at least $\frac{n^{1-\alpha}}{1-\alpha}$. Also,

$$
\begin{aligned}
(1 + \epsilon)^{2(1-\alpha)} n^{1-\alpha} \sum_{j \in \text{High}} (\hat{q}_j)^{1-\alpha} &= (1 + \epsilon)^{2(1-\alpha)} \left(\frac{n}{\ell}\right)^{1-\alpha} \sum_j (\hat{q}_j \ell)^{1-\alpha} \\
&= (1 + \epsilon)^{2(1-\alpha)} \left(\frac{n}{\ell}\right)^{1-\alpha} \text{cost}(\Pi, \ell)
\end{aligned}
$$

by definition of $\text{cost}(\Pi, \ell)$.

Therefore we showed,

$$
\begin{aligned}
\text{cost}(\Pi, n) &\leq \epsilon(1 + \epsilon)^{1-\alpha} \ell^{1-\alpha} \left(\frac{n}{\ell}\right)^{1-\alpha} + (1 + \epsilon)^{2(1-\alpha)} \left(\frac{n}{\ell}\right)^{1-\alpha} \text{cost}(\Pi, \ell) \\
&\leq (1 + 5\epsilon) \left(\frac{n}{\ell}\right)^{1-\alpha} \text{cost}(\Pi, \ell)
\end{aligned}
$$

The lower bound is basically similar. For $j \in Low$, we have $q_j < \left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}}$ and $\hat{q}_j <$

$\left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}} (1+\epsilon)$ (by Chernoff bounds). So we have

$$
\begin{aligned}
\sum_j (\hat{q}_j \ell)^{1-\alpha} &\leq \sum_j \left(\left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}} (1+\epsilon)\ell \right)^{1-\alpha} \\
&= r\frac{\epsilon}{r}(1+\epsilon)^{1-\alpha}\ell^{1-\alpha} \\
&= \epsilon(1+\epsilon)^{1-\alpha}n^{1-\alpha}\left(\frac{\ell}{n}\right)^{1-\alpha} \\
&\leq \epsilon(1+\epsilon)^{1-\alpha}\mathrm{cost}(\Pi,n)\left(\frac{\ell}{n}\right)^{1-\alpha}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathrm{cost}(\Pi,\ell) &= \sum_{j\in\mathrm{Low}} (\hat{q}_j\ell)^{1-\alpha} + \sum_{j\in\mathrm{High}} (\hat{q}_j\ell)^{1-\alpha} \\
&\leq \mathrm{cost}(\Pi,n)\epsilon\left(\frac{\ell}{n}\right)^{1-\alpha}(1+\epsilon)^{1-\alpha} + \sum_{j\in\mathrm{High}} (q_jn)^{1-\alpha}\left(\frac{\ell}{n}\right)^{1-\alpha}\left(\frac{\hat{q}_j}{q_j}\right)^{1-\alpha} \\
&\leq \mathrm{cost}(\Pi,n)\epsilon\left(\frac{\ell}{n}\right)^{1-\alpha}(1+\epsilon) + \sum_{j\in\mathrm{High}} (q_jn)^{1-\alpha}\left(\frac{\ell}{n}\right)^{1-\alpha}(1+\epsilon) \\
&\leq (1+\epsilon)^2\mathrm{cost}(\Pi,n)\left(\frac{\ell}{n}\right)^{1-\alpha}
\end{aligned}
$$

with probability at least $1 - \exp\{-q_jn\epsilon^2/2\} \geq 1 - \delta_3$. For low-probability j, the number of item j sold is $\geq \left(\frac{\epsilon}{r}\right)^{\frac{1}{1-\alpha}} n(1-\epsilon)$ with probability at least $1 - \delta_3$. A union bound extends these to all j with combined probability $1 - r\delta_3$.

Thus we obtain the upper bound: $\mathrm{cost}(\Pi,n) \leq \mathrm{cost}(\Pi,\ell)(1+\epsilon)^{2(1-\alpha)}\left(\frac{n}{\ell}\right)^{1-\alpha}$ and the lower bound: $\mathrm{cost}(\Pi,n) \geq \mathrm{cost}(\Pi,\ell)(1+\epsilon)^{-2}\left(\frac{n}{\ell}\right)^{1-\alpha}$, with probability at least $1 - r2^r(\delta_1 + \delta_2 + \delta_3)$.

A naive union bound can be done over all the permutations, which will add a factor of $r!$, we can reduce the factor to $r2^r$ by noticing that we are only interested in events of the type π_j, namely a given item (say, j) is in the set of desired items, and another set (say, $\{1,\ldots,j-1\}$) is not in that set. This has only $r2^r$ different events we need to perform the union over. $\qquad\square$

13.4.2 Generalized Performance Guarantees

We define $GreedyGeneralCost(\ell, n)$ as follows. For the first ℓ customers it allocates arbitrary items they desire, and observed their desired sets. Give the sets of the first ℓ customers, it runs $GreedyGeneralCost$ and computes a permutation $\hat{\Pi}$ of the items. For the remaining customers it allocates using permutation $\hat{\Pi}$. Namely, each customer is allocated the first item in the permutation $\hat{\Pi}$ that is in its desired set. The following theorem bounds the performance of $GreedyGeneralCost(\ell, n)$ for α-poly cost functions.

Theorem 13.11. *With probability $1 - \delta^{(\ell)}$ (for $\delta^{(\ell)}$ as in Theorem 13.10), the cost of* $GreedyGeneralCost(\ell, n)$ *is at most*

$$\ell + \frac{(1 + \epsilon)^{4-2\alpha}}{1 - \alpha} \mathrm{OPT}$$

Proof. Let $\hat{\Pi}$ be the permutation policy produced by GreedyGeneralCost, after the ℓ first customers. By Theorem 13.7,

$$\mathrm{cost}(\hat{\Pi}, \ell) \leq \frac{1}{1 - \alpha} \min_{\Pi} \mathrm{cost}(\Pi, \ell).$$

By Theorem 13.10, with probability $1 - \delta^{(\ell)}$,

$$\min_{\Pi} \mathrm{cost}(\Pi, \ell) \leq \min_{\Pi} \mathrm{cost}(\Pi, n)(1 + \epsilon)^2 \left(\frac{\ell}{n}\right)^{1-\alpha}.$$

Additionally, on this same event,

$$\mathrm{cost}(\hat{\Pi}, n) \leq \mathrm{cost}(\hat{\Pi}, \ell)(1 + \epsilon)^{2(1-\alpha)} \left(\frac{n}{\ell}\right)^{1-\alpha}.$$

Altogether, this implies

$$\mathrm{cost}(\hat{\Pi}, n) \leq \frac{(1 + \epsilon)^{2(1-\alpha)}}{1 - \alpha} \left(\frac{n}{\ell}\right)^{1-\alpha} \min_{\Pi} \mathrm{cost}(\Pi, n)(1 + \epsilon)^2 \left(\frac{\ell}{n}\right)^{1-\alpha}$$

$$= \frac{(1 + \epsilon)^{4-2\alpha}}{1 - \alpha} \min_{\Pi} \mathrm{cost}(\Pi, n).$$

$\square$

Corollary 13.12. *For any fixed constant $\delta \in (0, 1)$, for any*

$$\ell \geq \frac{3}{\epsilon^2} \left(\frac{r}{\epsilon}\right)^{\frac{1}{1-\alpha}} \ln\left(\frac{3r2^r}{\delta}\right),$$

and

$$n \geq \left(\frac{\ell}{\epsilon}\right)^{\frac{1}{1-\alpha}}$$

with probability at least $1 - \delta$ we have $GreedyGeneralCost(n, \ell)$ is at most

$$\left(\frac{(1+\epsilon)^{4-2\alpha}}{1-\alpha} + \epsilon\right) \text{OPT}$$

13.4.3 Generalization for β-nice costs

Toward extending the offline-model results under Assumption 13.6 to the online setting, consider the following lemma.

Lemma 13.13. *For any cost cost satisfying Assumption 13.6 with a given β, for any $k \geq 1$, the cost cost' with $\mathrm{cost}'_i(x) = \mathrm{cost}_i(kx)$ also satisfies Assumption 13.6 with the same β.*

Proof.

$$\frac{\mathrm{cost}_i(kx)}{x} = k\frac{\mathrm{cost}_i(kx)}{kx} \leq \beta k(\mathrm{cost}_i(kx) - \mathrm{cost}_i(kx - 1)).$$

Also, the property of nonincreasing marginal costs implies $\forall t \in \{1, \ldots, k\}$,

$$\mathrm{cost}_i(kx) - \mathrm{cost}_i(kx - 1) \leq \mathrm{cost}_i(kx - (t-1)) - \mathrm{cost}_i(kx - t),$$

so that

$$k(\mathrm{cost}_i(kx) - \mathrm{cost}_i(kx-1)) \leq \sum_{t=1}^{k}(\mathrm{cost}_i(kx-(t-1)) - \mathrm{cost}_i(kx-t)) = \mathrm{cost}_i(kx) - \mathrm{cost}_i(k(x-1)).$$

Therefore,

$$\frac{\mathrm{cost}_i(kx)}{x} \leq \beta(\mathrm{cost}_i(kx) - \mathrm{cost}_i(k(x-1))).$$

$\square$

Now the strategy is to run GreedyGeneralCost with the rescaled cost function $\text{cost}'_i(x) = \text{cost}_i(\frac{n}{\ell}x)$. This provides a β-approximation guarantee for the rescaled problem. The following theorem describes the generalization capabilities of this strategy.

Theorem 13.14. *Suppose $n \geq \ell$ and the cost function satisfies Assumption 13.6, and that $\forall i$, $\text{cost}_i(1) \in [1, B]$, where $B \geq 1$ is constant. Let $\text{cost}'_i(x) = \text{cost}_i(\frac{n}{\ell}x)$. With probability at least $1 - \delta^{(\ell)}$, for any permutations Π,*

$$\text{cost}'(\Pi, \ell)\frac{1-\epsilon}{1+2\epsilon-\epsilon^2} \leq \text{cost}(\Pi, n) \leq \text{cost}'(\Pi, \ell)\frac{(1+\epsilon)^2}{1-\epsilon},$$

where $\delta^{(\ell)} = r^2 2^{r+1}(\delta_1 + \delta_2)$, $\delta_1 = \exp\{-\epsilon^3 n^{\log_2(1+\frac{1}{2\beta})}/(3rB(1+\epsilon))\}$, and $\delta_2 = \exp\{-\epsilon^2\ell\frac{\epsilon}{rB(1+\epsilon)}n^{\log_2(1+\frac{1}{2\beta})-1}/3\}$. It is not necessary for the set of ℓ customers to be contained in the set of n customers for this.

Proof. Fix a permutation Π. Let π_j denote the event that a customer buys item Π_j and not covered by items Π_1 through Π_{j-1}. Namely, the probability that the consumer set of desired items include j and none of the items $1, \ldots, j-1$. Let q_j denote $Pr[\pi_j]$, and let $\hat{q}_j$ denote the fraction of Π_j on the initial ℓ-sample.

Let $q^* = \frac{\epsilon}{rB(1+\epsilon)}n^{c-1}$, where $c = \log_2(1+\frac{1}{2\beta})$. Item j is a "Low probability item" if $q_j < q^*$, and is called a "High probability item" if $q_j \geq q^*$. Let the set "Low" include all "Low probability items"; and the set "High" include all "High probability items".

First we address the case of item j of low probability. By a Chernoff bound, the quantity of item j that we will sell when applying Π to n customers is at most $q^*n(1+\epsilon)$, with probability at least $1 - \exp\{-\epsilon^2 q^*n/3\} = 1 - \delta_1$. By a union bound, this holds for all low probability items j with probability at least $1 - |\text{Low}|\delta_1$.

Next, suppose j has high probability. In this case, the quantity of item j we will sell when applying Π to n customers is at most $q_jn(1+\epsilon)$, with probability at least $1 - \exp\{-\epsilon^2 q_jn/3\} \geq 1 - \delta_1$. Again, a union bound implies this holds for all high probability j with probability at least $1 - |\text{High}|\delta_1$.

We have that (by Chernoff bounds), with probability at least $1 - \exp\{-\epsilon^2 \ell q_j / 3\} \geq 1 - \delta_2$, we have $q_j / \hat{q}_j \leq (1 + \epsilon)$. A union bound implies this holds for all high probability j with probability $1 - r\delta_2$.

Furthermore, noting that $q_j n (1 + \epsilon) = \hat{q}_j n (1 + \epsilon) \frac{q_j}{\hat{q}_j}$, and upper bounding $\frac{q_j}{\hat{q}_j}$ by $1 + \epsilon$, we get that $q_j n (1 + \epsilon) \leq (1 + \epsilon)^2 \hat{q}_j n$, with probability at least $1 - \delta_2$. Thus, with probability at least $1 - r\delta_1 - r\delta_2$,

$$
\begin{aligned}
\mathrm{cost}(\Pi, n) &\leq \mathrm{cost}(\mathbf{Low}) + \mathrm{cost}(\mathbf{High}) \\
&\leq \sum_{j \in \mathbf{Low}} \mathrm{cost}_j \left(q^* n (1 + \epsilon) \right) + \sum_{j \in \mathbf{High}} \mathrm{cost}_j \left((1 + \epsilon)^2 \hat{q}_j n \right) \\
&\leq r B q^* n (1 + \epsilon) + (1 + \epsilon)^2 \sum_{j \in \mathbf{High}} \mathrm{cost}_j \left(\hat{q}_j n \right) \\
&= r B q^* n (1 + \epsilon) + (1 + \epsilon)^2 \sum_{j \in \mathbf{High}} \mathrm{cost}'_j(\Pi, \ell).
\end{aligned}
$$

Note that Lemma 13.9 (with $\epsilon = 1/x$) implies that on n customers, $\mathrm{OPT} \geq \min_j \mathrm{cost}_j(n) \geq n^{\log_2(1 + \frac{1}{2\beta})} \min_j \mathrm{cost}_j(1) \geq n^{\log_2(1 + \frac{1}{2\beta})} = n^c$, where the third inequality is by the assumption on the range of $\mathrm{cost}_i(1)$. Thus, $r B q^* n (1 + \epsilon) = \epsilon n^c \leq \epsilon \mathrm{OPT}$.

We showed that

$$
\mathrm{cost}(\Pi, n) \qquad\qquad \leq \qquad\qquad \epsilon \mathrm{OPT} + (1 + \epsilon)^2 \sum_{j \in \mathbf{High}} \mathrm{cost}'_j(\Pi, \ell)
$$

$$
\leq \epsilon \mathrm{cost}(\Pi, n) + (1 + \epsilon)^2 \sum_{j \in \mathbf{High}} \mathrm{cost}'_j(\Pi, \ell).
$$

Therefore,

$$
\begin{aligned}
\mathrm{cost}(\Pi, n) &\leq \frac{(1 + \epsilon)^2}{1 - \epsilon} \sum_{j \in \mathbf{High}} \mathrm{cost}'_j(\Pi, \ell) \\
&\leq \frac{(1 + \epsilon)^2}{1 - \epsilon} \mathrm{cost}'(\Pi, \ell).
\end{aligned}
$$

The lower bound is basically similar. For $j \in \mathbf{Low}$, a Chernoff bound implies we have

$\hat{q}_j < q^*(1 + \epsilon)$ with probability at least $1 - \exp\{-\epsilon^2 q^* \ell / 3\} \geq 1 - \delta_2$. So we have

$$
\begin{aligned}
\sum_{j \in \text{Low}} \text{cost}_j(\hat{q}_j n) &\leq \sum_{j \in \text{Low}} \text{cost}_j(q^*(1 + \epsilon)n) \\
&\leq r B(1 + \epsilon) q^* n \\
&= \epsilon n^c \\
&\leq \epsilon \text{OPT} \\
&\leq \epsilon \text{cost}(\Pi, n).
\end{aligned}
$$

For $j \in$ High, again by a Chernoff bound, we have $\hat{q}_j / q_j \leq (1 + \epsilon)$ with probability at least $1 - \exp\{-\epsilon^2 q_j \ell / 3\} \geq 1 - \delta_2$. Thus, by a union bound, with probability at least $1 - r\delta_2$,

$$
\begin{aligned}
\text{cost}'(\Pi, \ell) &= \sum_{j \in \text{Low}} \text{cost}_j(\hat{q}_j n) + \sum_{j \in \text{High}} \text{cost}_j(\hat{q}_j n) \\
&\leq \epsilon \text{cost}(\Pi, n) + \sum_{j \in \text{High}} \text{cost}_j(q_j n(1 + \epsilon)).
\end{aligned}
$$

By another application of Chernoff and union bounds, with probability at least $1 - \sum_{j \in \text{High}} \exp\{-\epsilon^2 q_j n / 2\} \geq 1 - r\delta_1$, for every $j \in$ High, the number of j we will sell when applying Π to n customers is at least $q_i n(1 - \epsilon)$. Thus,

$$
\sum_{j \in \text{High}} \text{cost}_j(q_j n(1 + \epsilon)) = \sum_{j \in \text{High}} \text{cost}_j(q_j n(1 - \epsilon) \frac{1 + \epsilon}{1 - \epsilon}) \leq \frac{1 + \epsilon}{1 - \epsilon} \sum_{j \in \text{High}} \text{cost}_j(q_j n(1 - \epsilon)) \leq \frac{1 + \epsilon}{1 - \epsilon} \text{cost}(\Pi, n).
$$

Altogether, we have proven that with probability at least $1 - r(\delta_1 + \delta_2)$,

$$
\begin{aligned}
\text{cost}'(\Pi, \ell) &\leq \left(\epsilon + \frac{1 + \epsilon}{1 - \epsilon} \right) \text{cost}(\Pi, n) \\
&= \frac{1 + 2\epsilon - \epsilon^2}{1 - \epsilon} \text{cost}(\Pi, n),
\end{aligned}
$$

which implies

$$
\frac{1 - \epsilon}{1 + 2\epsilon - \epsilon^2} \text{cost}'(\Pi, \ell) \leq \text{cost}(\Pi, n).
$$

A naive union bound can be done over all the permutations, which will add a factor of $r!$; we can reduce the factor to $r2^r$ by noticing that we are only interested in events of the type π_j,

namely a given item (say, j) is in the set of desired items, and another set (say, $\{1, \ldots, j-1\}$) is not in that set. This has only $r2^r$ different events we need to perform the union over. Thus, the above inequalities hold for all permutations with probability at least $1 - r^2 2^{r+1}(\delta_1 + \delta_2)$. $\qquad \square$

Let $n_0 = 0$, $n_1 = 2\left(\frac{3rB(1+\epsilon)}{\epsilon^3} \ln\left(\frac{4r^2 2^{r+2}}{\delta}\right)\right)^{\frac{1}{\log_2(1+\frac{1}{2\beta})}}$. For each integer $i \geq 2$, define

$$n_i = \left(\frac{(\sum_{j=1}^{i-1} n_j)\epsilon^3}{3rB(1+\epsilon) \ln\left(\frac{(i+2)^2 r^2 2^{r+2}}{\delta}\right)}\right)^{\frac{1}{1-\log_2(1+\frac{1}{2\beta})}}.$$

We define $GreedyGeneralCost_\beta(n)$ as follows. Allocate arbitrary (valid) items to the first n_1 customers. For each $i \geq 2$ with $\sum_{j=1}^{i} n_i \leq n$, run $GreedyGeneralCost(\mathcal{S})$ with cost function cost', where $\mathcal{S}$ is the set of buyers $1, 2, \ldots, \sum_{j=1}^{i-1} n_j$, and $\forall j$, $\text{cost}'_j(x) = \text{cost}_j(xn_i/\sum_{t=1}^{i-1} n_t)$; this produces a permutation policy $\hat{\Pi}$. We then allocate to the customers $(\sum_{j=1}^{i-1} n_j)+1, \ldots, \sum_{j=1}^{i} n_j$ using the permutation policy $\hat{\Pi}$.

The following theorem bounds the performance of $GreedyGeneralCost_\beta(n)$.

Theorem 13.15. *If* cost *satisfies Assumption 13.6, and has* $\text{cost}_j(1) \in [1, B]$ *for every* $j \leq r$, *with probability at least* $1 - \delta$, *the cost of* $GreedyGeneralCost_\beta(n)$ *is at most*

$$Bn_1 + \beta \frac{(1+\epsilon)^2(1+2\epsilon - \epsilon^2)}{(1-\epsilon)^2} \sum_{i:\sum_{j=1}^{i} n_j \leq n} \text{OPT}(n_i).$$

Proof. By Theorem 13.7, Lemma 13.13, and Theorem 13.14 and a union bound, with probability at least $1-\delta$, for every i, the cost of $GreedyGeneralCost_\beta$ on customers $1+\sum_{j=1}^{i-1} n_j, \ldots, \sum_{j=1}^{i} n_j$ is at most

$$
\begin{aligned}
\text{cost}'\left(\hat{\Pi}, \sum_{j=1}^{i-1} n_j\right)\frac{(1+\epsilon)^2}{1-\epsilon} &\leq \beta \min_{\Pi} \text{cost}'\left(\Pi, \sum_{j=1}^{i-1} n_j\right)\frac{(1+\epsilon)^2}{1-\epsilon} \\
&\leq \beta \min_{\Pi} \text{cost}(\Pi, n_i)\frac{(1+\epsilon)^2(1+2\epsilon - \epsilon^2)}{(1-\epsilon)^2} \\
&= \beta \frac{(1+\epsilon)^2(1+2\epsilon - \epsilon^2)}{(1-\epsilon)^2}\text{OPT}(n_i).
\end{aligned}
$$

Summing over i yields the result. $\qquad \square$

302

If we are allowed to preview the utilities of some initial $o(n)$ set of buyers, then we can get the following simpler result.

Theorem 13.16. *If* cost *satisfies Assumption 13.6, and has* $\mathrm{cost}_j(1) \in [1, B]$ *for every* $j \leq r$, *with probability at least* $1 - \delta$, *the cost of applying the policy found by* $GreedyGeneralCost(\{1, \ldots, \ell\})$ *to all* n *customers is at most*

$$\beta \frac{(1+\epsilon)^2(1+2\epsilon-\epsilon^2)}{(1-\epsilon)^2}\mathrm{OPT}(n),$$

where $\ell = \left\lceil n^{1-\log_2(1+\frac{1}{2\beta})} \frac{3rB(1+\epsilon)}{\epsilon^3} \ln\left(\frac{r^2 2^{r+2}}{\delta}\right) \right\rceil = o(n)$.

Proof. By Theorem 13.7, Lemma 13.13, and Theorem 13.14, with probability at least $1 - \delta$, the cost of applying the policy $\hat{\Pi}$ found by $GreedyGeneralCost(\{1, \ldots, \ell\})$ to customers $1, \ldots, n$ is at most

$$
\begin{aligned}
\mathrm{cost}'(\hat{\Pi}, \ell)\frac{(1+\epsilon)^2}{1-\epsilon} &\leq \beta \min_{\Pi} \mathrm{cost}'(\Pi, \ell)\frac{(1+\epsilon)^2}{1-\epsilon} \\
&\leq \beta \min_{\Pi} \mathrm{cost}(\Pi, n)\frac{(1+\epsilon)^2(1+2\epsilon-\epsilon^2)}{(1-\epsilon)^2} \\
&= \beta \frac{(1+\epsilon)^2(1+2\epsilon-\epsilon^2)}{(1-\epsilon)^2}\mathrm{OPT}(n).
\end{aligned}
$$

$\square$

Also consider the following lemma.

Lemma 13.17. *If* cost *satisfies Assumption 13.6, then for any* $n \in \mathbb{N}$, $\mathrm{OPT}(2n) \geq \left(1 + \frac{1}{2\beta}\right)\mathrm{OPT}(n)$.

Proof. $\square$

We define $GreedyGeneralCost'_\beta(n)$ as follows. Allocate an arbitrary (valid) item to the first customer. For each $i \geq 1$ with $i \leq \log_2(n)$, run $GreedyGeneralCost(\mathcal{S})$, where $\mathcal{S}$ is the set of buyers $1, 2, \ldots, 2^{i-1}$; this produces a permutation policy $\hat{\Pi}$. We then allocate to the customers $2^{i-1} + 1, \ldots, 2^i$ using the permutation policy $\hat{\Pi}$.

The following theorem bounds the performance of $GreedyGeneralCost'_\beta(n)$.

303

Theorem 13.18. *If* cost *satisfies Assumption 13.6, and has* $\mathrm{cost}_j(1) \in [1, B]$ *for every* $j \leq r$, *letting* ℓ *denote the smallest power of* 2 *greater than* $\left(\frac{3rB(1+\epsilon)}{\epsilon^3} \ln \left(\frac{4r^2 2^{r+2}}{\delta} \right) \right)^{\frac{1}{\log_2(1 + \frac{1}{2\beta})}}$, *with probability at least* $1 - \sum_{i=\log_2(\ell)}^{\log_2(n)-1} r^2 2^{r+2} \left(\frac{\delta}{4r^2 2^{r+2}} \right)^{2^{(i-\log_2(\ell)) \log_2(1 + \frac{1}{2\beta})}}$, *the cost of* $GreedyGeneralCost'_\beta(n)$ *is at most*

$$B\ell + \frac{(1 + \epsilon)^2(1 + 2\epsilon - \epsilon^2)}{(1 - \epsilon)^2}(2\beta)^2 \mathrm{OPT}(n).$$

Proof. By Theorem 13.7, Theorem 13.14 and a union bound, with the stated probability, for every $i > \log_2(\ell)$, the cost of $GreedyGeneralCost'_\beta$ on customers $2^{i-1} + 1, \ldots, 2^i$ is at most

$$
\begin{aligned}
\mathrm{cost}\left(\hat{\Pi}, \{1, \ldots, 2^{i-1}\} \right) \frac{(1 + \epsilon)^2}{1 - \epsilon} &\leq \beta \min_{\Pi} \mathrm{cost}\left(\Pi, \{1, \ldots, 2^{i-1}\} \right) \frac{(1 + \epsilon)^2}{1 - \epsilon} \\
&\leq \beta \min_{\Pi} \mathrm{cost}(\Pi, \{2^{i-1} + 1, \ldots, 2^i\}) \frac{(1 + \epsilon)^2(1 + 2\epsilon - \epsilon^2)}{(1 - \epsilon)^2} \\
&= \beta \frac{(1 + \epsilon)^2(1 + 2\epsilon - \epsilon^2)}{(1 - \epsilon)^2} \mathrm{OPT}(\{2^{i-1} + 1, \ldots, 2^i\}).
\end{aligned}
$$

By Lemma 13.17,

$$\mathrm{OPT}(\{2^{i-1} + 1, \ldots, 2^i\}) = \mathrm{OPT}(2^{i-1}) \leq \mathrm{OPT}(2n2^{i-1-\lceil \log_2(n) \rceil})$$

$$\leq \left(\frac{1}{1 + \frac{1}{2\beta}} \right)^{\lceil \log_2(n) \rceil + 1 - i} 2\mathrm{OPT}(n).$$

Summing this over $i \in \{\log_2(\ell) + 1, \ldots, \lceil \log_2(n) \rceil\}$ is at most $4\beta \mathrm{OPT}(n)$. Plugging this into the above bound on the cost supplies the stated result. $\qquad\square$

13.5   General Unit Demand Utilities

In this section we show how to give a constant approximation for the case of general unit demand buyers in the offline setting in the case when we have a budget B to bound the cost we incur and we would like to maximize the buyers social welfare given this budget constraint. The main tool would be a reduction of our problem to the budgeted maximum coverage problem.

Definition 13.19. *An instance of the* budgeted maximum coverage problem *has a universe* X *of* m *elements where each* $x_i \in X$ *has an associated weight* w_i; *there is a collection of* m *sets*

$\mathcal{S}$ such that each sets $S_j \in \mathcal{S}$ has a cost c_j; and there is a budget L. A feasible solution is a collection of sets $\mathcal{S}' \subset \mathcal{S}$ such that $\sum_{S_j \in \mathcal{S}'} c_j \leq L$. The goal is to maximize the weight of the elements in $\mathcal{S}'$, i.e., $w(\mathcal{S}') = \sum_{x_i \in \cup_{S \in \mathcal{S}'} S} w_i$.

While the budgeted maximum coverage problem is NP-complete there is a $(1 - 1/e)$ approximation algorithm [Khuller, Moss, and Naor, 1999]. Their algorithm is a variation of the greedy algorithm, where on the one hand it computes the greedy allocation, where each time a set which maximizes the ratio between weight of the elements covered and the cost of the set is added, as long as the budget constraint is not violated. On the other hand the single best set is computed. The output is the best of the two alternative (either the single best set of the greedy allocation).

Before we show the reduction from a general unit demand utility to the budgeted maximum coverage problem, we show a simpler case where for each buyer j has a value v_j such that of any item i either $v_j = u_{j,i}$ or $u_{j,i} = 0$, which we call *buyer-uniform unit demand*.

Lemma 13.20. *There is a reduction from the budgeted buyer-uniform unit demand buyers problem to the budgeted maximum coverage problem. In addition the greedy algorithm can be computed in polynomial time on the resulting instance.*

Proof. For each buyer j we create an element x_j with weight v_j. For each item k and any subsets of buyers S we create a set $T_{S,k} = \{x_j : j \in S\}$ and has cost $cost_k(|S|)$. The budget is set to be $L = B$. Clearly any feasible allocation of the budgeted maximum coverage problem $T_{S_1,k_1}, \ldots T_{S_r,k_r}$ can be translated to a solution of the budgeted buyer-uniform unit demand buyers by simply producing item k_i for all the buyers in T_{S_i,k_i}. The welfare is the sum of the weight of the elements covered which is the social welfare, and the cost is exactly the production cost.

Note that the reduction generates an exponential number of sets, if we do it explicitly. However, we can run the Greedy algorithm easily, without generating the sets explicitly. Assume we have m' remaining buyers. For each item i and any $\ell \in [1, m']$ we compute the cost $cost_i(\ell)/gain_i(\ell)$, where $gain_i(\ell)$ is the weight of the ℓ buyers with highest valuation for item i. Greedy select the item i and number of buyers ℓ which have the highest ratio and adding this set

still satisfies the budget constraint. Note that given that greedy selects $T_{S,k}$ where $|S| = \ell$ then its cost is $cost_k(\ell)$ and its weigh is $w(T_{S,k}) \leq gain_k(\ell)$, and hence Greedy will always select one of the sets we are considering. □

In the above reduction we used very heavily the fact that each buyer j has a single valuation v_j regardless of which desired item it gets. In the following we show a slightly more involved reduction which handles the general unit demand buyers.

Lemma 13.21. *There is a reduction from the budgeted general unit demand buyers problem to the budgeted maximum coverage problem. In addition the greedy algorithm can be computed in polynomial time on the resulting instance.*

Proof. For each buyer j we sort its valuations $u_{j,i_1} \leq \cdots \leq u_{j,i_m}$. We set $v_{j,i_1} = u_{j,i_1}$ and $v_{j,i_r} = u_{j,i_r} - u_{j,i_{r-1}}$. Note that $\sum_{s=1}^{r} v_{j,i_s} = u_{j,i_r}$. For each buyer j we create m elements $x_{j,r}$, $1 \leq r \leq m$. For a buyer j and item k let $X_{j,k}$ be all the elements that represent lower valuation than $u_{j,k}$, i.e., $X_{j,k} = \{x_{j,r} : u_{j,i_r} \leq u_{j,k}\}$. For each item k and any subsets of buyers S we create a set $T_{S,k} = \cup_{j \in S} X_{j,k}$ and has cost $cost_k(|S|)$. The budget is set to be $L = B$.

Any feasible allocation of the budgeted maximum coverage problem $T_{S_1,k_1}, \ldots T_{S_l,k_r}$ can be translated to a solution of the budgeted general unit demand buyers producing item k_i for all the buyers in T_{S_i,k_i}. We call buyer j as *winner* if there exists some b such that $x_{j,b} \in \cup_{i=1}^{r} T_{S_i,k_i}$. Let $Winners$ we the set of all winner buyers. For any winner buyer $j \in Winner$ let $item(j) = s$ such that $s = \max\{b : x_{j,b} \in \cup_{i=1}^{r} T_{S_i,k_i}\}$.

The cost of our allocation is by definition at most $L = B$. The social welfare is

$$\sum_{x_{j,b} \in \cup_{i=1}^{r} T_{S_i,k_i}} v_{j,b} = \sum_{j \in Winner} u_{j,item(j)}$$

Again, note that the reduction generates an exponential number of sets, if we do it explicitly. However, we can run the Greedy algorithm easily, without generating the sets explicitly. For each item i and any $\ell \in [1, m]$ we compute the cost $cost_i(\ell)/gain_i(\ell)$, where $gain_i(\ell)$ is the weight of the ℓ buyers with highest valuation for item i. Greedy selects the item i and number

of buyers ℓ which have the highest ratio which still satisfies the budget constraint. Note that given that greedy selects $T_{S,k}$ where $|S| = \ell$ then its production cost is $cost_k(\ell)$ and its weight is $w(T_{S,k}) \leq gain_k(\ell)$, and hence Greedy will always select one of the sets we are considering. Once the Greedy selects a set $T_{S,k}$ we need to update the utility of any buyer $j \in S$ for any other item i, by setting $u_{j,i} = \max\{u_{j,i} - u_{j,k}, 0\}$, which is the residual valuation buyer j has for getting item i in addition to item k. $\qquad\square$

Combining our reduction with approximation algorithm of [Khuller, Moss, and Naor, 1999] we have the following theorem.

Theorem 13.22. *There exists a poly-time algorithm for the budgeted general unit demand buyers problem which achieves social welfare at least* $(1 - 1/e)\mathrm{OPT}$.

13.5.1 Generalization

To extend these results to the online setting, we will use Theorem 13.3 to represent allocations by pricing policies, and then use the results from above to learn a good pricing policy based on an initial sample.

Theorem 13.23. *Suppose every* $u_{j,i} \in [0, B]$. *With* $\ell = O((1/\epsilon^2)(r^3 \log(rB/\epsilon) + \log(1/\delta)))$ *random samples, with probability at least* $1 - \delta$, *the empirical per-customer social welfare is within* $\pm\epsilon$ *of the expected per-customer social welfare, uniformly over all price vectors in* $[0, B]^r$.

Proof. We will show that, for any distribution P and value $\epsilon > 0$, there exist $N = 2^{O(r^3 \log(rB/\epsilon))}$ functions $f_1, \ldots, f_N$ such that, for every price vector $\mathrm{price} \in [0, B]^r$, the function $g(x) = x_{\arg\max_{i \leq r} x_i - \mathrm{price}_i}$ has $\min_{k \leq N} \int |f_k - g| \mathrm{d}P \leq \epsilon$. This value N is known as the *uniform ϵ-covering number*. The result then follows from standard uniform convergence bounds (see e.g., [Haussler, 1992]).

The function $x \mapsto \max_{i \leq r} x_i - \mathrm{price}_i$ is a hyperplane with slope 1 in coordinate i and slope 0 in all other coordinates. So the subgraph (i.e., the set of $r + 1$-dimensional points (x, y) for

307

which $\max_{i \leq r} x_i - \text{price}_i \geq y$ is a union of r halfspaces in $r + 1$ dimensions. The space of unions of r halfspaces in $r + 1$ dimensions has VC dimension $r(r + 2)$, so this upper bounds the pseudo-dimension of the space of functions $\max_{i \leq r} x_i - \text{price}_i$, parametrized by the price vector price. Therefore, the uniform ϵ-covering number of this class is $2^{O(r^2 \log(B/\epsilon))}$.

For each $i \leq r$, the set of vectors $x \in [0, B]^r$ such that $i = \arg\max_k x_k - \text{price}_k$ is an intersection of r halfspaces in r dimensions. Thus, the function $x \mapsto \text{price}_{\arg\max_i x_i - \text{price}_i}$ is contained in the family of linear combinations of r disjoint intersections of r halfspaces. The VC dimension of an intersection of r halfspaces in r dimensions is $r(r + 1)$. So assuming the prices are bounded in a range $[0, B]$, the uniform ϵ-covering number for linear combinations (with weights in $[0, B]$) of r disjoint intersections of r halfspaces is $2^{O(r^3 \log(rB/\epsilon))}$. To prove this, we can take an $\epsilon/(2rB)$ cover (of $\{0, 1\}$-valued functions) of intersections of r halfspaces, which has size $(rB/\epsilon)^{O(r^2)}$, and then take an $\epsilon/(2r)$ grid in $[0, B]$ and multiply each function in the cover by each of these values to get a space of real-valued functions; there are $(rB/\epsilon)^{O(r^2)}$ total functions in this cover, and for each term in the linear combination of r disjoint intersections of r halfspaces, at least one of these real-valued functions will be within ϵ/r of it. Thus, taking the set of sums of r functions from this cover forms an ϵ-cover of the space of linear combinations of r disjoint intersections of r halfspaces, with size $(rB/\epsilon)^{O(r^3)}$.

Now note that $x_{\arg\max_i(x_i - \text{price}_i)} = \max_i(x_i - \text{price}_i) + \text{price}_{\arg\max_i(x_i - \text{price}_i)}$. So the uniform ϵ-covering number for the space of possible functions $x_{\arg\max_i(x_i - \text{price}_i)}$ is at most the produce of the uniform $(\epsilon/2)$-covering number for the space of functions $x \mapsto \max_i(x_i - \text{price}_i)$ and the uniform $(\epsilon/2)$-covering number for the space of functions $x \mapsto \text{price}_{\arg\max_i(x_i - \text{price}_i)}$; by the above, this produce is $2^{O(r^3 \log(rB/\epsilon))}$. $\qquad\square$

13.6   Properties of β-nice cost

Let $cost(n)$ be a β-nice cost function. We show a few properties of it.

Claim 13.24.

$$cost(2n) \geq cost(n) \left(1 + \frac{1}{2\beta}\right)$$

Proof. Let $a = cost(n)/n$ be the average cost of the first n items. Then the cost of the first $2n$ items is at least an, and has an average cost of at least $a/2$. The marginal cost of item $2n$ is at least $a/(2\beta)$. Therefore the cost of the items $n + 1$ to $2n$ is at least $an/(2\beta)$. □

We can get a better bound by a more refine analysis.

Claim 13.25. *Let $a_n = cost(n)/n$ be the average cost of the first n items. Then,*

$$a_{n+1} \geq a_n \frac{n}{n+1} \left(1 + \frac{1}{\beta(n+1)}\right)$$

and

$$a_n \geq a_1 \frac{1}{n} \prod_{t=1}^{n} \left(1 + \frac{1}{\beta(t+1)}\right) \geq e^{1/\beta^2} \cdot a_1 n^{-1+(1/\beta)}$$

Proof. The marginal cost of item $n+1$ is at least a_n/β. Therefore the cost of the first items $n+1$ is at least $na_n + a_n/(\beta)$, which gives the first expression.

We get the expression of a_n as a function of a_1 by repeatedly using the recursion. The approximation follows from,

$$\ln(a_n) \geq \ln(a_1) - \ln(n) + \sum_{t=1}^{n} \ln(1 + \frac{1}{\beta(n+1)})$$

$$\geq \ln(a_1) - \ln(n) + \sum_{t=1}^{n} \frac{1}{\beta(t+1)} - \frac{1}{(\beta(t+1))^2}$$

$$\geq \ln(a_1) - \ln(n) + \frac{1}{\beta}\ln(n) - \frac{1}{\beta^2}$$

where we used the identity $x - x^2 \leq \ln(1 + x)$. □

Bibliography

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. 12.5.2

N. Alon, B. Awerbuchy, Y. Azarz, N. Buchbinder, and J. Naor. The online set cover problem. *SIAM Journal on Computing*, 39(2):361–370, 2009. 13.1.2

R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. Technical Report RC23462, IBM T.J. Watson Research Center, 2004. 7.2.1

R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005. 7.2.1

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. 10.4.2, 10.4.2

M. A. Arcones. A Bernstein-type inequality for U-statistics and U-processes. *Statistics & Probability Letters*, 22(3):239 – 247, 1995. 2.3, 2.23

R. B. Ash and C. A. Doléans-Dade. *Probability & Measure Theory*. Academic Press, 2000. 6.2

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007. 12.1.1, 12.2.2

P. Awasthi, M.-F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. *arXiv:1307.8371v2*, 2013. 11.5, 11.5, 11.11, 11.5, 11.5, 11.14, 11.5

M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the* 26$^{\text{th}}$ *Conference on Learning Theory*, 2013. 11.12

M. F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006. 2.1, 12.4

M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT)*, 2007a. 2.1, 6.1

M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the* 20$^{\text{th}}$ *Conference on Learning Theory*, 2007b. 11.5

M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *Proceedings of the 21st Annual Conference on Computational Learning Theory (COLT)*, 2008. 2.1

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009. 6.1, 12.4

M.-F. Balcan, S. Hanneke, and J. W. Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2–3):111–139, September 2010. 6.1, 6.3, 6.5, 7.4, 10.7, 12.5.2, 12.5.2

Z. Bar-Yossef. Sampling lower bounds via information theory. In *Proceedings of the 35th Annual ACM Symposium on the Theory of Computing*, pages 335–344, 2003. 8.4

P. Bartlett, M. I. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006. 12.1, 12.1.1, 12.2.1, 12.2.1, 12.2.1, 12.2.2, 12.2.2, 12.2.2, 12.2.2, 12.2.2, 12.2.4, 12.3.2, 12.4, 12.5.1, 12.5.1, 12.5.1, 12.5.1, 12.5.1, 12.5.5

P. L. Bartlett. Learning with a slowly changing distribution. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 243–252, 1992. 10.1, 10.3

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(11):463–482, 2002. 12.4

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. 12.1, 12.2.4

R. D. Barve and P. M. Long. On the complexity of learning from drifting distributions. *Inf. Comput.*, 138(2):170–193, 1997. 10.3

E. Baum and K. Lang. Query learning can work poorly when a human oracle is used. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, 1993. 2.1

J. Baxter. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28:7–39, 1997. 7.1, 7.2.1, 7.4, 8.1

J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12: 149–198, 2000. 7.2.1

M. Bellare, O. Goldreich, and M. Sudan. Free bits, PCPs and non-approximability – towards tight results. *SIAM J. Comput.*, 27(3):804–915, 1998. 2.1.1

S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *Conference on Learning Theory*, 2003. 7.1, 7.2.1

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009. 2.1, 6.1, 12.1.1, 12.3.2, 12.4, 12.5.2

E. Blais. Testing juntas nearly optimally. In *Proc. 41st Annual ACM Symposium on the Theory of Computing*, pages 151–158, 2009. 2.1

A. Blum, A. Gupta, Y. Mansour, and A. Sharma. Welfare and profit maximization with production costs. In *FOCS*, pages 77–86, 2011. 13.1.2

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. 8.3, 10.8

N. H. Bshouty, Y. Li, and P. M. Long. Using the doubling dimension to analyze the generalization

of learning algorithms. *Journal of Computer and System Sciences*, 75(6):323–335, 2009. ISSN 0022-0000. 5.2

J. G. Carbonell. Learning by analogy: Formulating and generalizing plans from past experience. In R. S. Michalski, J. G.Carbonell, and T. M. Mitchell, editors, *Machine Learning, An Artificial Intelligence Approach*. Tioga Press, Palo Alto, CA, 1983. 7.1

J. G. Carbonell. Derivational analogy: A theory of reconstructive problem solving and expertise acquisition. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning, An Artificial Intelligence Approach, Volume II*. Morgan Kaufmann, 1986. 7.1

R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997. 7.1

R. Castro and R. Nowak. Minimax bounds for active learning. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT)*, 2007. 2.1

R. Castro and R. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, July 2008. 6.1, 12.5.4

O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. MIT press, 2006. 2.1.2, 2.4, 2.11, 2.11, 8

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 201–221, 1994a. 2.1

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994b. 10.1, 10.4

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. 12.2.2

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2006. 5.1, 5.4, 5.4

P. Cramton, Y. Shoham, and R. Steinberg. *Combinatorial Auctions*. The MIT Press, 2006. 9.1, 9.3

S. Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems*, pages 337–344. MIT Press, 2004. 6.1

S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*, volume 18, 2005. 2.1, 5.5, 6.1, 6.2, 6.3, 10.7

S. Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 2011. To appear. 2.1.2, 2.4

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. Technical Report CS2007-0898, Department of Computer Science and Engineering, University of California, San Diego, 2007a. 10.5.2, 12.4, 12.5.2

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. *Advances in Neural Information Processing Systems*, 20, 2007b. 2.1, 6.1

S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, 10:281–299, 2009. 6.1, 10.3, 11.5

J. V. Davis and I. Dhillon. Differential entropic clustering of multivariate gaussians. In *Advances in Neural Information Processing Systems 19*, 2006. 2.10.4

O. Dekel, C. Gentile, and K. Sridharam. Robust selective sampling from single and multiple teachers. In *Conference on Learning Theory*, 2010. 10.3

N. R. Devanur and T. P. Hayes. The adwords problem: Online keyword matching with budgeted bidders under random permutations. In *Proc. ACM EC*, EC '09, pages 71–78, 2009. 13.1.2

N. R. Devanur and K. Jain. Online matching with concave returns. In *Proc. STOC*, pages 137–144, 2012. 13.1.2

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, NY, USA, 2001. 7.2, 7.5, 7.3, 8.3, 8.4, 9.2

I. Diakonikolas, H. Lee, K. Matulef, K. Onak, R. Rubinfeld, R. Servedio, and A. Wan. Testing for concise representations. In *Proc. 48th Annual IEEE Symposium on Foundations of Computer*

Science, pages 549–558, 2007. 2.1

E. Dolev and D. Ron. Distribution-free testing algorithms for monomials with a sublinear number of queries. In *Proceedings of the 13th international conference on Approximation, and 14 the International conference on Randomization, and combinatorial optimization: algorithms and techniques*, APPROX/RANDOM'10, pages 531–544. Springer-Verlag, 2010. 2.1.1

R. M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 15(4): 1306–1326, 1987. 12.5.4

T. Evgeniou and M. Pontil. Regularized multi-task learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004. 7.1, 7.2.1

T. Evgeniou, C. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005. 7.2.1

E. Fischer. The art of uninformed decisions. *Bulletin of the EATCS*, 75:97–126, 2001. 2.6

E. Fischer, G. Kindler, D. Ron, S. Safra, and A. Samorodnitsky. Testing juntas. *J. Comput. Syst. Sci.*, 68:753–787, 2004. 2.1

Y. Freund and Y. Mansour. Learning under persistent drift. In *Proceedings of the Third European Conference on Computational Learning Theory*, EuroCOLT '97, pages 109–118, 1997. ISBN 3-540-62685-9. 10.3

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. 12.2.2

Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2):133–168, 1997. 5.5, 6.1, 6.2

E. Friedman. Active learning for smooth problems. In *Proceedings of the 22nd Conference on Learning Theory*, 2009. 6.1, 12.5.2, 12.5.2

A. Gersho. Asymptotically optimal block quantization. *IEEE Transactions on Information The-*

ory, 25(4):373–380, 1979. 5.1

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006. 12.2.4, 12.2.4, 12.3.1, 12.4, 12.5.2, 12.5.3, 12.5.4, 12.5.4, 12.5.4, 12.5.6

E. Giné, V. Koltchinskii, and J. Wellner. Ratio limit theorems for empirical processes. In *Stochastic Inequalities*, pages 249–278. Birkhäuser, 2003. 12.5.3

D. Glasner and R. A. Servedio. Distribution-free testing lower bound for basic boolean functions. *Theory of Computing*, 5(1):191–216, 2009. 2.1.1

G. Goel and A. Mehta. Online budgeted matching in random input models with applications to adwords. In *Proc. SODA*, pages 982–991, 2008. 13.1.2

O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998. 2.1

A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *In Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, 2003. 5.2, 5.2, 5.4, 5.7

S. Halevy and E. Kushilevitz. Distribution-free connectivity testing. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 3122 of *Lecture Notes in Computer Science*, pages 393–404. Springer Berlin / Heidelberg, 2004. 2.1.1

S. Halevy and E. Kushilevitz. A lower bound for distribution-free monotonicity testing. In *Approximation, Randomization and Combinatorial Optimization*, volume 3624 of *Lecture Notes in Computer Science*, pages 612–612. Springer Berlin / Heidelberg, 2005. 2.1.1

S. Halevy and E. Kushilevitz. Distribution-free property-testing. *SIAM Journal on Computing*, 37(4):1107–1138, 2007. 2.1.1, 3

S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th Annual International Conference on Machine Learning (ICML)*, 2007a. 2.1, 5.5, 6.1,

10.3, 12.5.2, 12.5.2

S. Hanneke. Teaching dimension and the complexity of active learning. In *In Proceedings of the* *20ᵗʰ Annual Conference on Learning Theory*, 2007b. 5.5, 6.1

S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009. 6.1, 6.5, 6.3, 6.6, 7.1, 7.4, 12.1, 12.5.2

S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011. 6.1, 10.3, 10.5.2, 10.5.3, 10.8, 10.9, 10.18, 10.10, 12.1, 12.5.1, 12.5.2, 12.5.2, 12.5.4, 12.5.4

S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13:1469–1587, 2012. 12.4, 12.5.1, 12.5.2, 12.5.2

S. Hanneke and L. Yang. Negative results for active learning with convex losses. In *Proceedings of the* 13ᵗʰ *International Conference on Artificial Intelligence and Statistics*, 2010. 12.3.2

D. Haussler. Decision theoretic generalizations of the PAC model for neural net a nd other learning applications. *Information and Computation*, 100:78–150, 1992. 9.1, 12.5.4, 13.5.1

D. Haussler, M. Kearns, and R. Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning*, 14(1):83–113, 1994a. 5.5, 6.1, 6.2, 6.5

D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115:248–292, 1994b. 10.4, 10.8

N. J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008. ISBN 978-0-898716-46-7. 2.10.4

D. A. Huffman. A method for the construction of minimum-redundancy codes. In *Proceedings of the I.R.E.*, pages 1098–1102, 1952. 5.4

M. Kääriäinen. Active learning in the non-realizable case. In *In Proc. of the 17th International*

Conference on Algorithmic Learning Theory, 2006. 6.1

O. Kallenberg. *Foundations of Modern Probability, 2nd Edition*. Springer Verlag, New York, 2002. 7.3.1

R. M. Karp, U. V. Vazirani, and V. V. Vazirani. An optimal algorithm for on-line bipartite matching. In *Proc. STOC*, pages 352–358, 1990. 13.1.2

M. Kearns. *The Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989. 4.2

M. Kearns and D. Ron. Testing problems with sublearning sample complexity. *Journal of Computer and System Sciences*, 61(3):428 – 456, 2000. 2, 2.1, 2.1.2, 2.2

M. Kearns, M. Li, and L. Valiant. Learning boolean formulas. *J. ACM*, 41:1298–1328, November 1994. 4.2

S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Inf. Process. Lett.*, 70(1):39–45, 1999. 13.1.1, 13.5, 13.5

J. C. Kieffer. A survey of the theory of source coding with a fidelity criterion. *IEEE Transactions on Information Theory*, 39(5):1473–1490, 1993. 5.1, 5.3

E. E.-D. Koby Crammer, Yishay Mansour and J. W. Vaughan. Regret minimization with concept drift. In *COLT*, pages 168–180, 2010. 10.1, 11.1, 11.3, 11.5, 11.5

J. Kolodner (Ed). *Case-Based Learning*. Kluwer Academic Publishers, The Netherlands, 1993. 7.1

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001. 12.4

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. 10.9, 12.1, 12.2.4, 12.2.4, 12.3.1, 12.4, 12.5.1, 12.5.1

V. Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems: Lecture notes. Technical report, Ecole d'ete de Probabilités de Saint-Flour, 2008. 12.6

V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning Research*, To Appear, 2010. 6.1, 12.1, 12.1.1, 12.3.2, 12.4, 12.5.2

S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11(1):23–35, 1993. 5.5

L. Li, M. L. Littman, and T. J. Walsh. Knows what it knows: A framework for self-aware learning. In *International Conference on Machine Learning*, 2008. 10.6

S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011. 11.5

T. Linder and R. Zamir. One the asymptotic tightness of the shannon lower bound. *IEEE Transactions on Information Theory*, 40(6):2026–2031, 1994. 5.1

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988. 10.3

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. 2.1.2

S. Mahalanabis. A note on active learning for smooth problems. arXiv*:1103.3095*, 2011. 12.5.2, 12.5.2

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27: 1808–1829, 1999. 10.1, 12.2.4, 12.5.1, 12.5.1

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *In Advances in Neural Information Processing Systems (NIPS)*, pages 1041–1048, 2008. 10.3

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009. 10.3

K. Matulef, R. O'Donnell, R. Rubinfeld, and R. A. Servedio. Testing halfspaces. In *Proc. 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 256–264, 2009. 2.1, 2.1.2, 2.1.2, 2.3, 2.10, 2.3

A. Mehta, A. Saberi, U. Vazirani, and V. Vazirani. Adwords and generalized online matching. *J. ACM*, 54(5), 2007. 13.1.2

C. Micchelli and M. Pontil. Kernels for multi–task learning. In *Advances in Neural Information Processing 18*, 2004. 7.1, 7.2.1

S. Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13(1): 67–90, 2012. 12.1.1

N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007. ISBN 0521872820. 13.1.2, 13.3.2

D. Nolan and D. Pollard. U-processes: Rates of convergence. *The Annals of Statistics*, 15(2): 780–799, 1987. 12.5.4

R. D. Nowak. Generalized binary search. In *Proceedings of the 46$^{\text{th}}$ Annual Allerton Conference on Communication, Control, and Computing*, 2008. 6.1

M. Parnas, D. Ron, and A. Samorodnitsky. Testing basic boolean formulae. *SIAM J. Discret. Math.*, 16(1):20–46, 2003. 2.1.1

J. Poland and M. Hutter. MDL convergence speed for Bernoulli sequences. *Statistics and Computing*, 16:161–175, 2006. 8.4

D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, Berlin / New York, 1984. 12.5.4

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. 12.5.4

E. C. Posner and E. R. Rodemich. Epsilon entropy and data compression. *The Annals of Mathe-*

matical Statistics, 42(6):2079–2125, 1971. 5.1

E. C. Posner, E. R. Rodemich, and H. Rumsey, Jr. Epsilon entropy of stochastic processes. *The Annals of Mathematical Statistics*, 38(4):1000–1020, 1967. 5.1

M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems 24*, 2011. 12.5.2, 12.5.2, 12.5.4, 12.5.4

D. Ron. Property testing: A learning theory perspective. *Foundations and Trends in Machine Learning*, 1(3):307–402, 2008. 2.1

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000. 8

R. Rubinfeld and M. Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM J. Comput.*, 25:252–271, 1996. 2.1.1

R. E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5:197–227, July 1990. 4.2

M. J. Schervish. *Theory of Statistics*. Springer, New York, NY, USA, 1995. 7.2, 7.3, 8.2, 8.4, 9.2

R. A. Servedio. On learning monotone DNF under product distributions. *Information and Computation*, 193:57–74, 2004. 4.4, 4.4

H. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the 5th Annual ACM workshop on Computational learning theory*, pages 287–294, 1992. 2.1

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423,623–656, 1948. 5.1, 5.3, 5.4

C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Rec., Part 4*, pages 142–163, 1959. 5.1

G. E. Shilov. *Linear Algebra*. Dover, 1977. 2.10.4

D. L. Silver. *Selective Transfer of Neural Network Task Knowledge*. PhD thesis, Computer Science, University of Western Ontario, 2000. 7.1

J. Tenenbaum, V. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000. 8

S. Thrun. Is learning the n-th thing any easier than learning the first? In *In Advances in Neural Information Processing Systems 8*, 1996. 7.1

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 4:45–66, 2001. 2.1

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. 12.2.4, 12.5.1, 12.5.1

S. van de Geer. *Empirical Processes in M-Estimation (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, 2000a. 10.3

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000b. 9.2

A. van der Vaart and J. A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011. 12.2, 12.5.3, 12.5.3

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996. 12.2.4, 12.5.3, 12.5.3, 12.5.4, 12.5.5, 12.5.6, 12.6

V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982. 7.2, 7.3, 7.4.1, 8.2, 8.3, 9.2, 10.4.2, 10.4.2, 10.8

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. 11.3, 12.5.4

V. N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998. 2.11

M. M. Veloso and J. G. Carbonell. Derivational analogy in prodigy: Automating case acquisition, storage and utilization. *Machine Learning*, 10:249–278, 1993. 7.1

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applica-*

tions, chapter 5, pages 210–268. Cambridge University Press, 2012. Available at `http://arxiv.org/abs/1011.3027`. 2.5.2, 2.27, 2.10.4

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117–186, 1945. 8.4

L. Wang. Sufficient conditions for agnostic active learnable. In *Advances in Neural Information Processing Systems 22*, 2009. 6.1

L. Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12:2269–2292, 2011. 12.5.2, 12.5.2

L. Yang, S. Hanneke, and J. Carbonell. Identifiability of priors from bounded sample sizes with applications to transfer learning. In *24*th *Annual Conference on Learning Theory*, 2011. 8.1, 8.2, 8.3

L. Yang, S. Hanneke, and J. Carbonell. A theory of transfer learning with applications to active learning. *Machine Learning*, 90(2):161–189, 2013. 9, 9.1, 9.2, 9.2

Y. G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13:768–774, 1985. 7.5, 7.3, 8.3, 9.2

P. L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–149, 1982. 5.1

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004. 12.1.1

M. Zinkevich, A. Blum, and T. Sandholm. On polynomial-time preference elicitation with value queries. In *Proceedings of the* 4th *ACM Conference on Electronic Commerce*, pages 175–185, 2003. 9.1, 9.3