Active Learning with Noise

Aaron Geelon So*

September 7, 2019

Abstract

This is an expository paper on a general framework for the design and analysis of active learning algorithms. The usual passive formulation of machine learning assumes that training data is collected i.i.d. from the world, without influence from the learner. Active learning is a setting where the learner is able to collect its own training data—the hope is that the active learner can collect highly informative data to reduce the amount of data required to learn.

The main contribution of this thesis is in how it frames active learning algorithms: use the geometric structure of the model to deduce which statistics are especially informative with respect to the goal of learning, then actively collect data to estimate these statistics. In addition to aiding the design of active learning algorithms, this framework also enables us to conceptually distinguish the parametric nature of the model from the non-parametric nature of noise in our analysis. Naturally, the amount of data required to learn will depend on both the complexity of the model and the amount of noise/uncertainty in the data; we discuss common techniques to lower bound the amount of data required by active learning also.

*Columbia University, Master's thesis in Computer Science, geelon.so@columbia.edu

Contents

1 Introduction

Consider the supervised learning problem of binary classification [BBL05]. In the usual formulation, *instances* x from an *instance space* $\mathcal{X}$ are associated with classes y from a *label space* $\mathcal{Y}$. Their 'real-world' relationship is formally encapsulated by some unknown joint distribution P_{XY} over $\mathcal{X} \times \mathcal{Y}$. The learner, on the other hand, has an *a priori* model $\mathcal{H}$ for how instances may be associated with labels. Namely, a model or *hypothesis class* is a collection of functions or *hypotheses* of the form $h : \mathcal{X} \to \mathcal{Y}$.

To decide which hypothesis best captures the relationship between $\mathcal{X}$ and $\mathcal{Y}$, the learner relies on data $(x_1, y_1), \ldots, (x_n, y_n)$ collected from the real world—perhaps they are drawn from P_{XY}. It is likely the case that drawing a single sample will reveal relatively little about P_{XY}, but as more samples are drawn, more information about P_{XY} is learned. With enough samples, at some point, it will become possible to satisfactorily fit the model to the world.

The usual *passive* formulation of machine learning assumes that the samples are drawn i.i.d. from the underlying distribution. It is passive because the learner has no influence over what data it sees. But as not all data are equally useful, a learner that may *actively* collect data—perform experiments—may be able to learn the same as its passive counterpart, but using much less data. These notes are concerned with the amount of data n that enables an active learner to learn well.

Acknowledgments

I am very grateful to my advisor Daniel Hsu for introducing me to this field and for patiently providing me with invaluable and careful guidance throughout the whole research process. I am also very thankful for the constant support from Nakul Verma and for helping me deepen my understanding of machine learning. Thank you also to Chris Tosh for helpful discussion and suggestions for this thesis.

This work is also possible only due to the copious amount of existing work in machine learning and active learning. I have attempted to document the most immediately relevant references, but I am indebted to a much larger collection of works, and any lack of reference is not a claim of originality (besides the presentation and minor extensions of previously known results).

2 Learning as an estimation problem

The framework in which we've just described learning comes from classical statistical learning theory [BBL03], where learning is ultimately an *optimization* problem: find the hypothesis $h \in \mathcal{H}$ that is least likely to err. We might, for instance, optimize over the expected rate of misclassification as a notion of *error* or *risk*:

$$R(P_{XY}, h) := \mathop{\mathbb{E}}_{(x,y)\sim P_{XY}} \big[\mathbf{1}\{h(x) \neq y\}\big].$$

That is, we wish to obtain the *best-in-class classifier* h^*,

$$h^* := \arg\min_{h\in\mathcal{H}} R(h).$$

Note that for brevity, we omit P_{XY} and let $R(h) \equiv R(P_{XY}, h)$ when it is clear from context.

To solve this optimization problem, though, we can hardly expect to access P_{XY} as a closed analytic expression; more likely, we access P_{XY} through samples (x, y) drawn from it. One reasonable approach to learning is to use samples drawn from P_{XY} to estimate the risk—if the estimator is a sufficiently faithful proxy to the true risk, minimizing the estimated risk will approximately minimize the true risk. Thus, the problem of learning is transformed into one of *estimation*.

2.1 Empirical risk minimization

To see how learning through statistical estimation plays out, let's consider one very natural approach to estimating risk: draw n i.i.d. samples $(X_i, Y_i) \sim P_{XY}$ and construct the *empirical risk* of h on n samples:

$$\widehat{R}(h) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{h(X_i) \neq Y_i\}.$$

The law of large numbers tells us that $\widehat{R}(h)$ is an unbiased and consistent estimator of $R(h)$; that is, not only is the estimator correct in expectation, $\mathbb{E}\big[\widehat{R}(h)\big] = R(h)$, but also following intuition, we obtain increasingly higher quality estimates of the true risk through greater amounts of data.

In fact, large deviation bounds allow us to statistically quantify this trade off between quality of estimation and sample size. Consider the following upper bound on the probability that an empirical estimator fails to be close to its expected value:

Theorem 1 (Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_n$ be n i.i.d. Bernoulli variables. Then:*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z]\right| > \varepsilon\right] \leq 2\exp\left(-2n\varepsilon^2\right).$$

It immediately follows that if we wish to estimate the risk of a fixed hypothesis to an accuracy of $\pm\varepsilon$ with failure probability δ (i.e. the empirical risk fails to be within $\pm\varepsilon$ of the true risk with probability at most δ), we need to set the right-hand side of the inequality to δ and solve for n. This shows us that the empirical estimator constructed on n i.i.d. samples suffices, where the number of samples n needs be no larger than:

$$n = O\left(\frac{1}{\varepsilon^2}\ln\frac{1}{\delta}\right).$$

Now, this is the number of samples to estimate the risk of a single fixed hypothesis to within $\pm\varepsilon$. But in order to minimize risk over the whole hypothesis class $\mathcal{H}$, we'd need to estimate the risk for each hypothesis to within $\pm\varepsilon$. For simplicity, assume that $\mathcal{H}$ is finite; then we can use those n samples to construct $|\mathcal{H}|$ estimators in parallel. By a union bound, the probability that at least one of the estimators fails to fall within $\pm\varepsilon$ of its expected value is $|\mathcal{H}| \cdot \delta$.

Reparametrizing δ, we can now say that if we wish to simultaneously estimate the risk of all hypotheses of a finite hypothesis class to an accuracy of $\pm\varepsilon$ with failure probability δ (i.e. the probability that even one of the empirical risks fails to be within $\pm\varepsilon$ of its true risk is at most δ), it suffices to construct our $|\mathcal{H}|$ estimators using n i.i.d. samples. The number of samples we need to learn—the *sample complexity*—needs be no larger than:

$$n = O\left(\frac{1}{\varepsilon^2}\ln\frac{|\mathcal{H}|}{\delta}\right). \tag{1}$$

Notice that once we can guarantee that the empirical risks of each hypothesis in $\mathcal{H}$ is within $\pm\varepsilon$ of its true risk, then it must be the case that the *empirical risk minimizer* $h_{\text{ERM}} := \arg\min_{h\in\mathcal{H}} \widehat{R}(h)$ must satisfy:

$$R(h_{\text{ERM}}) \leq R(h^*) + 2\varepsilon.$$

In other words, the estimated hypothesis h_{ERM} is worse than the best-in-class hypothesis h^* by an error rate of at most 2ε with probability $1-\delta$, thus (ε, δ)-learning $\mathcal{H}$ (see Appendix A for a formal definition).

2.2 Hypothesis classes as pseudometric spaces

The core question posed by statistical learning—using data to recover $h^* \in \mathcal{H}$ up to some error tolerance—can in some sense, be formulated geometrically: a learner just has to return any hypothesis $\widehat{h}$ from some "ε-ball around h^*". But in what sense—with respect to which notion of distance?[1] We can define the distance between two hypotheses h and h' by the probability that they disagree on $\mathcal{X}$, where each instance in $\mathcal{X}$ is weighted by some marginal distribution P_X (usually, with respect to the true data distribution):

Definition 2 (Induced pseudometric on $\mathcal{H}$)**.** *Let P_X be a marginal distribution on $\mathcal{X}$. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a hypothesis class. The* pseudometric induced on $\mathcal{H}$ *by P_X is the map $d : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, defined by:*

$$d(h, h') := \Pr_{X \sim P_X} \big[h(X) \neq h'(X)\big].$$

To convince you that d is a proper pseudometric, we can rewrite it more familiarly as the L^1-metric:

$$d(h, h') \equiv \|h - h'\|_1 = \int_{\mathcal{X}} |h(x) - h'(x)| \, dP_X(x).$$

When $d(h, h') = 0$, then h and h' are equal almost everywhere. In fact, the L^1 interpretation of distance can be extended to help us describe the error of a hypothesis. Denote by g the conditional expectation:

$$g(x) := \mathbb{E}[Y|X = x].$$

If we view $\mathcal{H} \subset L^1(\mathcal{X}; [0,1])$ as a subset of L^1-integrable functions from $\mathcal{X}$ to $[0,1]$, then the error of a hypothesis h can then be rewritten as $R(h) = \|h - g\|_1$. And so, we could write $R(h) = d(h, g)$.

With this understanding of distances between hypotheses, we can characterize ε-learning as the geometric problem of searching for a hypothesis from within a ball around h^*, up to the *model error* $\nu := R(h^*)$. And in fact, if our model perfectly fit the the underlying distribution (i.e. $R(h^*) = 0$), then learning is exactly equivalent to the problem of search. Consider the following two lemmas, following from the triangle inequality. The first states that if we can search for an ε-close hypothesis to h^*, we can ε-learn $\mathcal{H}$. The second states that if we can ε-learn, we can search for a $(2\nu + \varepsilon)$-close hypothesis to h^*.

Lemma 3 (Search implies learn)**.** *Let $\mathcal{H}$ and P_{XY} be a hypothesis class and a joint probability distribution over $\mathcal{X} \times \mathcal{Y}$, as above. Let d be the pseudometric induced on $\mathcal{H}$ by the marginal distribution P_X. If $h \in \mathcal{H}$ satisfies $d(h, h^*) < \varepsilon$, its excess risk is at most ε.*

Lemma 4 (Learn implies search)**.** *Let $\mathcal{H}$, P_{XY}, and d as before. Let ν be the optimal risk of any $h \in \mathcal{H}$; that is, ν is the true risk $R(h^*)$ of a best-in-class classifier $h^* \in \mathcal{H}$:*

$$\nu := \min_{h \in \mathcal{H}} R(h).$$

If $h \in \mathcal{H}$ has excess risk at most ε, then $d(h, h^*) < 2\nu + \varepsilon$.

As a prelude to the next section in which we formulate learning as a search problem, let's reconsider how it is that empirical risk minimization (ε, δ)-learns $\mathcal{H}$:

1. Observe that the risk functional $R : \mathcal{H} \to \mathbb{R}$ with respect to P_{XY} contains enough information to identify h^* from $\mathcal{H}$; indeed h^* is the minimizer of R.

[1] We'll actually use a relaxed notion of distance/metric where we don't require distinct points to have nonzero distance. In particular, a *pseudometric* on a space X is a non-negative function $d : X \times X \to \mathbb{R}$ such that (i) $d(x, x) = 0$, (ii) $d(x, y) = d(y, x)$, and (iii) $d(x, y) + d(y, z) \geq d(x, z)$. The third condition is the usual triangle-inequality.

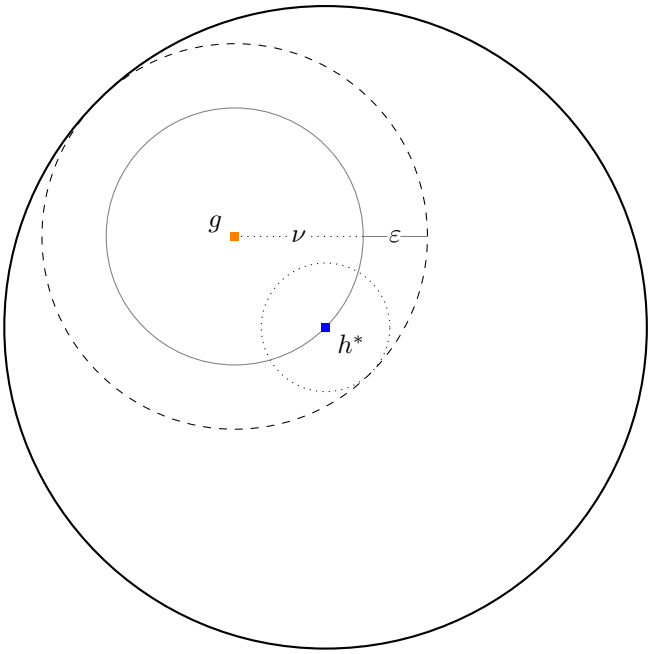

Figure 1: A pictorial proof of Lemmas 3 and 4. View $g(x) := \mathbb{E}[Y|X = x]$ and $h^* \in \mathcal{H}$ as elements in a pseudometric space. Here, $d(g, h^*) = \nu$. To ε-learn $\mathcal{H}$ means to return any hypothesis in the dashed circle centered at g. To ε- and $(2\nu + \varepsilon)$-search means to return any hypothesis in the dotted and solid circles centered at h^*, respectively.

2. Given an estimate $\widehat{R} : \mathcal{H} \to \mathbb{R}$ that is within $(\varepsilon/2)$-close of R with respect to the L^∞-norm:

$$\left\|\widehat{R} - R\right\|_\infty \leq \frac{\varepsilon}{2},$$

   it becomes possible to compute a set of hypotheses $h \in \mathcal{H}$ within ε of h^*, $S \subset B(h^*, 2\nu + \varepsilon)$.

3. The above analysis that combines Hoeffding's inequality with a union bound shows how such an estimator can be constructed using finite samples from P_{XY} with high probability.

In short, if we weakened the problem of ε-learning $\mathcal{H}$ to searching for a $(2\nu + \varepsilon)$-close hypothesis to h^* within $\mathcal{H}$, the overall technique to learning is simply to construct a single statistic (here, the estimator $\widehat{R}$) that carries enough information to reduce the search space $\mathcal{H}$ down to a small subset around the target h^*.

But of course, this statistic $\widehat{R}$ contains much more information than needed to convey the contents of some small non-empty subset of $B(h^*, 2\nu + \varepsilon)$; it requires so much information from labeled data that it is capable of estimating the risk for every hypothesis in $\mathcal{H}$ to within $\pm\varepsilon$. Might it be possible to construct other statistics or even *sequences* of statistics that still enable us to recover an appropriate subset $\varnothing \subsetneq S \subset B(h^*, 2\nu + \varepsilon)$, but require much less labeled data to estimate? In particular, perhaps it is possible to construct a sequence of statistics, each with lower information content—thus requiring less data to estimate—but together progressively reduces the search space. Before we investigate this idea formally, let's build up our intuition for how we can attack this problem with a simple thought experiment.

2.3 To the booksellers

Imagine walking into a bookstore. Though you've never been before, know that it's filled with countless volumes ranging from Greek tragedies to quantum mechanics, architectural coffee-table books to bathroom

readers. Out of millions of books, you wish to find one: Feller's *An Introduction to Probability Theory*.

Let's assume that the books are organized broadly by category and perhaps more finely in alphabetical order by author—and let's say that this model is a reasonable one, though perhaps a book might be misplaced here and there due to a careless patron.

One way you could attempt to locate this book is to first try to estimate the bookstore's organizational system: randomly pull books off the shelves and make a note of its category and author. If you do this for long enough, you'll obtain a pretty fine estimation, and this will allow you to jump into the introductory probability section of the bookstore and pull a book by an author whose last name is close to Feller's. Maybe you obtain Billingsley's *Probability and Measure*, and maybe that's ε-close enough for you.

But let's not use this is a ridiculous algorithm to search for a book in a bookstore. Instead, you might be more naturally inclined to browse around to build a high-level estimation of how the bookstore is organized. Once you've discovered the math/stats section, you probably want to concentrate your search in that area and estimate how their subfields are organized. Finally, at some point, you might even perform a binary search on the author's last name.

Even if the books aren't perfectly ordered, this latter algorithm could yield immense savings in how many books you need to look at. Here, the savings are due to being able to adaptively change where you look for the book. First, this allows you to decide not to gather information from areas that are likely irrelevant to your search (despite the off-chance that, say, someone moved Feller's book into the literary criticism section). In particular, because you can restrict your search space, you need to look at much fewer books before you get to the author-level precision. Stated differently, if adaptation weren't allowed or you hadn't restricted your search space over time, then to be able to identify an introductory probability book at the author level, you would've had to gather enough information to identify books from all fields at the author level; in such a case, perhaps the first algorithm (i.e. empirical risk maximization) is optimal. And second, adaptation allows you, when provided certain structural assumptions like the alphabetical ordering of books, to significantly speed up finding the book by seeking out maximally informative data points.

When adaptation is possible, this imaginary bookstore illustrates how we might try to learn using much less data. In the rest of the paper, we'll formalize these intuitions into techniques and analyses for learning with less data. Notice though, that if the bookstore is in disarray with books reshelved willy-nilly (i.e. noise or model misspecification), then perhaps no amount of adaptation could allow us to locate the book faster than essentially performing a brute-force search. And so, in addition to techniques and analyses, we will also study the information-theoretic lower bounds to understand what statistical limits exist for learning.

3 Learning as a search problem

By approaching learning from a search perspective, we implicitly mean that the learner may sequentially interact with the world: the knowledge it accumulates interacting with the world informs it on how to gather the next bit of information. We call such a learner an *active learner* because it actively chooses the information it uses to learn; or, in other words, the learner performs *experimental design*, for in choosing the next piece of information, it is simply choosing the next experiment to run.

Let's formalize a simple model of active learning: as before, we let $\mathcal{X}, \mathcal{Y}, \mathcal{H}$ be the instance, label, and hypothesis space. Let $\mathcal{P}$ be the class of possible 'real-world' data distributions, and let $P_{XY} \in \mathcal{P}$ is the underlying distribution over $\mathcal{X} \times \mathcal{Y}$. When the learner draws a sample $(x, y) \sim P_{XY}$, the label y is initially hidden. At a cost, the learner may ask a LABEL oracle to reveal the hidden label (e.g. the oracle could be a human providing feedback, the result of a physical experiment, so on). Thus, we're interested in minimizing the amount of calls to the LABEL oracle—the *label complexity*—required to learn well.

3.1 A mellow active learner: don't label uninformative points

Recall back to our booksellers example: as we learn more about the layout of the bookstore, we can concentrate our data collection efforts to areas we think are relevant. In other words, we might expect the size of the search space to decrease over time as we learn more about the bookstore. Indeed, we could perform the following natural algorithm: iterate between (i) exploring the current search space to learn about the bookstore's organization based on our prior model of a bookstore, and (ii) reducing the search space through our deepened understanding of how the bookstore is organized in the real world.

In the context of learning, let's say that at time t, having performed t experiments, we know that the best-in-class model $h^* \in \mathcal{H}$ is contained within a small subset of hypotheses we call a *version space*:

$$h^* \in \mathcal{V}_t \subset \mathcal{H}.$$

Then, a coarse proxy for how quickly our 'understanding of the real world deepens' could be how quickly these version spaces shrink. In particular, if each of these version spaces are contained within a ball:

$$\mathcal{V}_t \subset B(h^*, r_t),$$

we'd hope that the radius r_t decreases quickly as a function of t. And if we know that h^* is contained within $\mathcal{V}_t$, we shouldn't bother to label points $x \in \mathcal{X}$ on which all hypotheses $h \in \mathcal{V}_t$ agree: they will all be correct or incorrect together. So, let us define the *disagreement region* $\text{DIS}(\mathcal{V}_t) \subset \mathcal{X}$ to be the points x for which there is disagreement among $h \in \mathcal{V}_t$ as to what its label is:

$$\text{DIS}(\mathcal{V}_t) := \{x \in \mathcal{X} : \exists h, h' \in \mathcal{H} \text{ s.t. } h(x) \neq h'(x)\}.$$

The size of the disagreement region of $\mathcal{V}$ is just its probability mass, $P_X[\text{DIS}(\mathcal{V})]$. We could formalize in the following way the relationship between increasing knowledge of h^* and reducing the area of data collection in terms of shrinking the radius of a ball $B(h^*, r)$ containing $\mathcal{V}$ and decreasing the probability mass of the disagreement region $P_X[\text{DIS}(\mathcal{V})]$:

Definition 5. *The* disagreement coefficient θ_s *at resolution* s *with respect to* h^* *is the quantity:*

$$\theta_s(\mathcal{H}, P_X, h^*) := \sup_{r > s} \frac{P_X\left[\text{DIS}(B(h^*, r))\right]}{r}.$$

Simply by rearranging, we immediately obtain the following lemma from the definition:

Lemma 6. *Let* $\mathcal{V}_t \subset B(h^*, r)$ *be a version space of hypotheses* r*-close to* h^**, where* $r > s$*. Then, the disagreement region of* $\mathcal{V}_t$ *has probability mass at most* $\theta_s r$*.*

Notice that Lemma 4 and Lemma 6 together give us precisely the ingredients to recast our natural bookstore search algorithm: iterate between (i) reducing the size of our version space by estimating risk with respect to its disagreement region, as described by Lemma 4, and (ii) reducing the size of the disagreement region by shrinking the version space, as described by Lemma 6.

To see why reducing the size of the disagreement region helps us learn with fewer labels, recall the goal of learning: to obtain a hypothesis $\widehat{h} \in \mathcal{H}$ such that the excess risk is bounded by ε:

$$R(\widehat{h}) - R(h^*) \leq \varepsilon.$$

Usually, we can achieve this by empirically estimating the risk of each hypothesis up to $\pm\varepsilon/2$. But suppose that we have a version space $\mathcal{V} \subset \mathcal{H}$ that contains h^* and whose disagreement region has mass $\mu := P_X[\text{DIS}(\mathcal{V})]$. Then, we claim that we only need to estimate the error of each hypothesis in $\mathcal{V}$ up to $\pm\varepsilon/2\mu$ over $\text{DIS}(\mathcal{V})$.

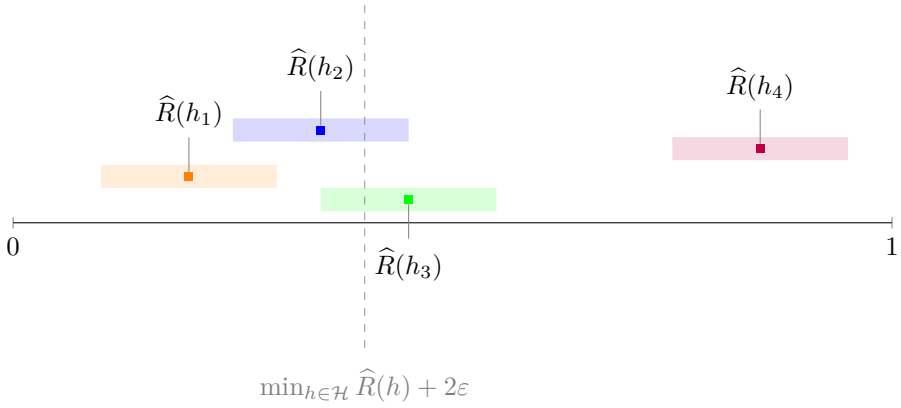

Figure 2: Let $\mathcal{V}_t = \{h_1, h_2, h_3, h_4\}$ be a version space. Drawing data from $\mathrm{DIS}(\mathcal{V}_t)$ allows us to generate estimates of risk, up to some $\pm\varepsilon$ precision, indicated by the colored bands around the estimated risks. In other words, the true risks for each of the hypotheses must fall within their respective colored bands. Since $\widehat{R}(h^*) \leq \min_{h\in\mathcal{H}} \widehat{R}(h) + 2\varepsilon$, the best-in-class hypothesis h^* must be to the left of the dotted line: either h_1 or h_2. We can now restrict our next version space $\mathcal{V}_{t+1} = \{h_1, h_2\}$ to just these two hypotheses.

Indeed, the error rate of $\widehat{h}$ and h^* are certainly equal on a $(1-\mu)$-fraction of $\mathcal{X}$, so points in the agreement region of $\mathcal{V}$ contribute nothing to the difference $R(\widehat{h}) - R(h^*)$; we just need to estimate the excess risk on the remaining μ-fraction of $\mathcal{X}$. Let's denote by $R\big|_{\mathrm{DIS}(\mathcal{V})}(h)$ to be the risk of h with respect to the truncated distribution that results from drawing $(x, y) \sim P_{XY}$ conditional on $x \in \mathrm{DIS}(\mathcal{V})$. We have:

$$R(h) - R(h^*) = \mu \cdot \left[R\big|_{\mathrm{DIS}(\mathcal{V})}(h) - R\big|_{\mathrm{DIS}(\mathcal{V})}(h^*) \right].$$

It follows that we just need to empirically estimate the risk with respect to the truncated distribution $R\big|_{\mathrm{DIS}(\mathcal{V})}$ up to $\pm\varepsilon/2\mu$ for all $h \in \mathcal{V}$. The same sample complexity bounds from before applies, so we need at most n data points to perform this estimation, where

$$n = O\left(\frac{\mu^2}{\varepsilon^2} \ln \frac{|\mathcal{V}|}{\delta} \right). \tag{2}$$

This statement about probability corresponds directly to our earlier discussion on searching for books: conceptually, ε parametrizes the precision to which we wish to be able to locate books—say, to the author level. When we can restrict the size of the search area within the store, we need to look at much fewer books before we attain that author-level precision. We can further translate our intuitive bookstore search algorithm to the following active learning algorithm, which is called A^2, for agnostic active learning, in the literature.

We'll iterate over T rounds between (i) estimating the risk of hypotheses in the current version space $\mathcal{V}_t$ using data $\mathcal{Z}_t$ drawn from its disagreement region according to its corresponding truncated distribution $P_{X|X\in\mathrm{DIS}(\mathcal{V}_t)}$, and (ii) paring down the version space by discarding hypotheses whose empirical risk estimates $\widehat{R}_{\mathcal{Z}_t}$ using the samples $\mathcal{Z}_t$ are far away from the minimum empirical risk R_t, thus further reducing the search space. In the following algorithm, k_t denotes the sample size used to estimate risk and ε_t is the 'far away' threshold we use to shrink the version space—we'll set their values later.

Algorithm A^2
($*$ disagreement-based learning $*$)
Input: k_t a sequence of positive integers, ε_t a sequence of positive reals
Initialize: $\mathcal{V}_0 \leftarrow \mathcal{H}$
1. **for** rounds $t = 0$ **to** T
2. **do** draw and label k_t data points, $\mathcal{Z}_t = \left\{\left(x_i^{(t)}, y_i^{(t)}\right) : 1 \le i \le k_t\right\}$
3. where $x_i^{(t)} \sim P_{X|X\in\mathrm{DIS}(\mathcal{V}_t)}$ and $y_i^{(t)} = \textsc{label}\left(x_i^{(t)}\right)$
4. compute the empirical risks $\widehat{R}_{\mathcal{Z}_t}(h) \leftarrow \frac{1}{k_t}\sum_{i=1}^{k_t} \mathbf{1}\{h(x) \neq y\}$ for all $h \in \mathcal{V}_t$
5. $R_t \leftarrow \min_{h\in\mathcal{V}_t} \widehat{R}_{\mathcal{Z}_t}(h)$
6. $\mathcal{V}_{t+1} \leftarrow \{h \in \mathcal{V}_t : \widehat{R}_{\mathcal{Z}_t}(h) \le R_t + 2\varepsilon_t\}$
7. **return** $\arg\min_{h\in\mathcal{V}_T} \widehat{R}_{\mathcal{Z}_T}(h)$

Suppose at time t, we've estimated the empirical risks of hypotheses in $\mathcal{V}_t$ to an ε_t-precision. As illustrated in Figure 2, we can safely discard any hypothesis whose empirical risk is more than $2\varepsilon_t$ the minimum empirical risk without discarding h^*. Having discarded those hypotheses with empirical risk more than $2\varepsilon_t$, we're left with a version space $\mathcal{V}_{t+1}$ whose hypotheses have excess risks of no more than $4\varepsilon_t$. According to Lemma 4, $\mathcal{V}_{t+1}$ fits into a ball $B(h^*, 2\nu + 4\varepsilon_t)$. Not only does the version space shrink, but so does the size μ_t of the disagreement region:

$$\mu_t = P_X[\mathrm{DIS}(\mathcal{V}_{t+1})] \le \theta(2\nu + 4\varepsilon_t).$$

Let's choose ε_t so that we iteratively halve our uncertainty: $\varepsilon_{t+1} = \varepsilon_t/2$. Our above discussion shows that we need to estimate the risk with respect to the disagreement region up to $\pm\varepsilon_t/2\mu_t$, requiring k_t labels, where:[2]

$$k_t = \widetilde{O}\left(\frac{\theta^2(2\nu + 4\varepsilon_t)^2}{(\varepsilon_t/2)^2}\log\frac{|\mathcal{V}|}{\delta}\right). \tag{3}$$

If at each step we halve ε_t, we'll need at most $T = \log\frac{1}{\varepsilon}$ rounds before we reach the desired accuracy, $\varepsilon_T \le \varepsilon$. And so, we deduce the following label complexity bound on A^2:

Theorem 7 (A^2 label complexity)**.** *Let $\varepsilon, \delta > 0$ be fixed. Let $\theta := \theta_{\nu+\varepsilon}(\mathcal{H}, P_X, h^*)$. There exists settings for k_t and ε_t such that the A^2 algorithm (ε, δ)-learns $\mathcal{H}$ with label complexity $n = \sum_{i=1}^{T} k_t$ at most:*

$$n = O\left(\theta^2\left(1 + \frac{\nu^2}{\varepsilon^2}\right)\log\frac{1}{\varepsilon}\log\frac{|\mathcal{H}|\log\frac{1}{\varepsilon}}{\delta}\right).$$

[2]Though setting k_t to this value is sufficient to perform the estimation for one round, technically, we would want to set k_t in the algorithm to

$$k_t = O\left(\frac{\theta^2(\nu+\varepsilon)^2}{\varepsilon^2}\log\frac{|\mathcal{H}|\log\frac{1}{\varepsilon}}{\delta}\right),$$

since we will end up performing $\log\frac{1}{\varepsilon}$ rounds of estimation. We'll need to estimate at most $|\mathcal{H}|\log\frac{1}{\varepsilon}$ quantities over the T rounds, so to keep the overall failure probability less than δ, we ensure that the failure probability for any single estimation is less than $\delta/|\mathcal{H}|\log\frac{1}{\varepsilon}$. And as $\log\log\frac{1}{\varepsilon}$ is very small, we use the $\widetilde{O}(\cdot)$ notation to suppress polylogarithmic terms in $\log\frac{1}{\varepsilon}$.

3.1.1 Comparison with ERM

To get a sense for how this label complexity can be much better than the usual label complexity for ERM, suppose we have a hypothesis class with disagreement coefficient $\theta = O(1)$ that is constant with respect to ε. Further, if there exists a hypothesis in $\mathcal{H}$ with very low true risk, $\nu \ll \varepsilon$, then the A^2 label complexity is exponentially better than the label complexity of ERM in the same setting:

$$n_{A^2} = \widetilde{O}\left(\log\frac{1}{\varepsilon}\log\frac{|\mathcal{H}|}{\delta}\right) \quad \text{vs.} \quad n_{\text{ERM}} = O\left(\frac{1}{\varepsilon}\log\frac{|\mathcal{H}|}{\delta}\right).$$

The bookstore analogy may once again be a useful to build up intuition for the setting with constant disagreement coefficient and low true risk, and why such an improvement in label complexity is possible. As before, we have a prior model $\mathcal{H}$ for how bookstores can organize books. We can think of the pseudometric on $\mathcal{H}$ as a notion of how similar two organizational systems are. If the disagreement coefficient is small, this means that a good understanding of how a bookstore is organized, $\mathcal{V} \subset B(h^*, \varepsilon)$, translates into the ability to locate many individual books quickly, $P_X[\text{DIS}(\mathcal{V})] \ll 1$. If ν is small, then our model fits the real-world well; in that case, we might also be able to iteratively halve our search space by inspecting a constant k_t number of books each time.

On the other hand, because $\theta_r \leq 1/r$ for all $r > 0$, in the worst case scenario, their label complexities are:

$$n_{A^2} = \widetilde{O}\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\log\frac{|\mathcal{H}|}{\delta}\right) \quad \text{vs.} \quad n_{\text{ERM}} = O\left(\frac{1}{\varepsilon^2}\log\frac{|\mathcal{H}|}{\delta}\right),$$

where A^2 has slightly worse label complexity due to having performed about $\log\frac{1}{\varepsilon}$ essentially independent estimation problems that did not contribute to simplifying the next estimation problem.

So far, both A^2 and ERM use data in order to estimate statistics about risk functionals. Recall that we formulated empirical risk minimization earlier as a one-shot estimation problem: through estimating the risk functional R to a high enough accuracy, we obtain sufficient information to find a hypothesis h such that $R(h) \leq R(h^*) + \varepsilon$. The A^2 algorithm simply does this in multiple rounds: over T rounds, construct and estimate the sequence of risk functionals $R\big|_{\text{DIS}(\mathcal{V}_0)}, \ldots, R\big|_{\text{DIS}(\mathcal{V}_T)}$. The analysis we perform just ensures that together, these statistics carry enough information to recover some h with excess risk bounded above by ε.

One natural question that follows is whether we can learn via other procedures besides directly estimating the optimization parameter of statistical learning (i.e. the risk functionals)? In particular, can we gather data to compute other statistics through which we can still recover an ε-good classifier? As a simple example, performing a binary search on the author's name makes use of data in a different way: it pares down the version space by utilizing the model's structural assumptions (e.g. books are ordered alphabetically by the author's last name). In this next section, we'll investigate how we might leverage the 'geometry' or 'shape' of the hypothesis class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ to learn using less data.

3.2 An aggressive active learner: label the maximally informative points

In the previous section, we used the disagreement coefficient to link the amount of information known about the hypothesis to the amount of information known about individual points: a small radius of $\mathcal{V} \subset B(h^*, r)$ translates to a large, what we might call, agreement region. Or so we hope. In this section, we aim to study the other direction: from individual data points, what can we deduce about the overall hypothesis? Are there points that give more useful information than other points? And so, while in the previous section, we sped up learning by querying points with nonzero information content, here, we'll attempt to maximize the amount of information we learn every time we call the LABEL oracle. This marks the passage from learning via non-parametric techniques to parametric ones.

To elaborate, we have so far used data to estimate risk functionals, which is a non-parametric task; while this task has been specified over a parametric class of hypotheses $\mathcal{H}$, we made use of the parameterization insofar as to convert the problem of estimating the absolute risks of the hypotheses to the weaker problem of estimating their relative risks. But we now enter the parametric regime as we aim to infer from the structure of a hypothesis class itself which data points are most informative. The assumptions that can make parametric tasks much easier than their non-parametric counterparts also present a new kind of problem: what happens if the real world does not conform to those assumptions?

As a simple example, let's consider the hypothesis class $\mathcal{H} = \{\mathbf{0}, \mathbf{1}\}$, which has perhaps the most structure of all: either everything is labeled 0 or everything is labeled 1. This parametric class of hypotheses makes extremely stringent assumptions about the underlying data distribution. If those assumptions are correct, then a single labeled example is enough to learn $\mathcal{H}$. But if not, or as we relax the class $\mathcal{P}$ of possible data distributions—as $R(h^*)$ approaches $\frac{1}{2}$—this parametric problem re-enters the non-parametric regime: there really is no way to learn but to estimate $\mathbb{E}_{P_{XY}}[Y]$, potentially up to $\pm\varepsilon$.

This tells us that our goal should not be to fully solve the learning problem in a parametric way, for if we allow for noise and model misspecification, there will come a point when the non-parametric aspect to our problem takes over; indeed, we will investigate this further when we study the lower bounds of active learning. But for now, let us recall Lemma 4, which showed that an ε-learner must be able to solve the weaker $(2\nu + \varepsilon)$-search problem. Of course, ERM can achieve this with label complexity:

$$O\left(\frac{1}{(2\nu+\varepsilon)^2}\log\frac{|\mathcal{H}|}{\delta}\right).$$

We will set for ourselves the modest goal of reducing the label complexity to get to this point: finding a version space $\mathcal{V} \subset \mathcal{H}$ that fits into the $(2\nu + \varepsilon)$-ball around h^*. If we impose no further condition on P_{XY} besides $R(h^*) = \nu$, to aim to reduce the radius beyond $2\nu + \varepsilon$ to less than ε could mean that (i) we attempt to solve a strictly harder problem than ε-learning, as shown in Lemma 3, and (ii) we enter the non-parametric regime, for if we have two hypotheses h, h' that disagree on at most a 2ν-fraction of points in $\mathcal{X}$ (i.e. $d(h, h') \leq 2\nu$) with the assumption that the noise rate is ν, the noise rate on their disagreement region could approach arbitrarily close to $\frac{1}{2}$, returning us to above situation with $\mathcal{H} = \{\mathbf{0}, \mathbf{1}\}$. Simply put, binary search would be useless if books were not actually in alphabetical order, or at least nearly so. We set the goal to reduce the version space to fit into the $(2\nu + \varepsilon)$-ball around h^* because, here, it is still the case that the data distribution nearly satisfies the parametric assumptions. So as to not overload ε, let's introduce a new parameter r, which you can feel free to read as $2\nu + \varepsilon$ (though we'll keep it free for generality).

To this end, let us view a version space $\mathcal{V}$ as a graph—the hypotheses in $\mathcal{V}$ are vertices, and the length of an edge $\{h, h'\}$ is the distance $d(h, h')$. Define the *diameter* of a set to be the maximal distance between any two points in the set:

$$\operatorname{diam}(\mathcal{V}) := \max_{h,h' \in \mathcal{V}} d(h, h').$$

If we can construct a version space $\mathcal{V}$ that contains h^* with diameter at most r, then $\mathcal{V}$ will certainly fit into $B(h^*, r)$. We'll quantify informativeness with respect to reducing the diameter of $\mathcal{V}$.

Consider how one usually goes about conducting an experiment: (i) collect data z from the world, then (ii) use data to compute some test statistics t_z, which is used to test alternative hypotheses[3]. Should we return to the bookstore, over to the statistics section, we could hypothesize that Feller's book is contained in the first half. As a possible test of this hypothesis, we could 'collect data' by taking a look at who wrote the book in the middle of the shelf: if the author's name comes after Feller's, then we know that our hypothesis

[3]Here, by *hypothesis*, we don't mean some $h \in \mathcal{H}$, but some postulated truth about the world. As an example, given a subset $S \subset \mathcal{H}$, a hypothesis H could be the statement 'h^* *is contained in* S'.

was correct. This appears to be a particularly informative experiment to perform, for no matter the outcome, we'll have obtained enough information to halve our search space.

Our goal now will be to work up to a functional notion of an 'informative test' with respect to our learning problem. But to do so, we'll need to formalize what it means to collect data and to compute a test. In the following, if $Z_1, \ldots, Z_m$ is a sequence of random variables, we'll refer to that sequence as Z^m for brevity.

Definition 8 (Learning strategy)**.** *Let $\mathcal{X}$ be the instance space and $\mathcal{Y}$ the label space. An m-step* learning strategy S *is a collection of m conditional probability distribution $\Pi^{(i)} \equiv \Pi^{(i)}_{X_i|X^{i-1},Y^{i-1}}$:*

$$\Pi^{(i)}_{X_i|X^{i-1},Y^{i-1}}\left(\cdot\,|x^{i-1}, y^{i-1}\right),$$

where $1 \leq i \leq m$. A learning strategy is passive *if for all i, $\Pi^{(i)} \equiv P_X$ is the marginal distribution on $\mathcal{X}$.*

We'll sequentially draw data from these distributions: once we've obtained the first $(i-1)$ data points, $(x_1, y_1), \ldots, (x_{i-1}, y_{i-1})$, we obtain the ith instance x_i according to the above ith distribution. Then, we label x_i by querying the LABEL oracle, which returns $y_i \sim P_{Y|X=x_i}$, a conditional distribution of the underlying joint data distribution P_{XY}. Let $\mathbb{P}^{(\mathsf{S}_m, P_{Y|X})}$ denote the distribution on the m-sample data set $(x_1, y_1), \ldots, (x_m, y_m)$ obtained in this way. For short, we'll refer to this as $\mathbb{P}^{\mathsf{S}_m}$. We can now define a test:

Definition 9 (Version space test)**.** *Let $\mathcal{V}$ be a version space. A* version space test *T using m samples is a random variable on $z^m \sim \mathbb{P}^{\mathsf{S}_m}$, where the outcome is a functional $t_{z^m} : \mathcal{V} \to \{0, 1\}$.*

Put simply, the outcome of a test assigns every hypothesis $h \in \mathcal{V}$ to one of two groups based on a sample z^m drawn from the sampling distribution $\mathbb{P}^{\mathsf{S}_m}$. In particular, the outcome t of a test splits a version space $\mathcal{V}$ into two subsets: $\mathcal{V}^0_t = \{h \in \mathcal{V} : t(h) = 0\}$, and $\mathcal{V}^1_t$, defined analogously. As a simple example of a test, fix some $x \in \mathcal{X}$. Define the test T_x that obtains a label $y \leftarrow \textsc{label}(x)$, and partitions $\mathcal{H}$ into those that made a mistake on x, and those that didn't. Formally, if δ_x is the Dirac distribution at x, define the sampling distribution to be $\mathbb{P}^{\mathsf{S}_1} := \delta_x \otimes P_{Y|X=x}$. Data collection just corresponds to getting x labeled. The test we described is the random variable $t_{(x,y)}(h) := \mathbf{1}\{h(x) \neq y\}$.

Now that we've formalized the notion of a test, let's consider what an informative test in the context of our search problem: intuitively, for a given version space, an informative test should be able to significantly reduce the diameter of the search space containing h^*. We can do so by comparing the number of edges that remain in the induced subgraphs on $\mathcal{V}^0_t$ and $\mathcal{V}^1_t$ with respect to the original version space $\mathcal{V}$. And as we care about being able to locate h^*, the test T should map h^* into $\mathcal{V}^0_t$ with positive advantage Δ:

$$\Pr_{z \sim \mathbb{P}^{\mathsf{S}}}\left[t_z(h^*) = 0\right] > \frac{1}{2} + \Delta.$$

We can think of $\mathcal{V}^1_t$ as a set of hypotheses that the test t indicates are outside the ball $B(h^*, r)$.

One way an active learner can more aggressively learn is by only performing experiments that are guaranteed to split off a large fraction of edges. To formalize this, let $Q \subset \binom{\mathcal{H}}{2}$ be any collection of edges. If T is a test, we say that it *ρ-splits Q with advantage Δ* if any outcome t removes a ρ-fraction of edges,

$$\left|Q \cap \binom{\mathcal{V}^0_t}{2}\right| \leq (1-\rho)|Q|,$$

and with probability at least $\frac{1}{2} + \Delta$, if $h^* \in \mathcal{V}$, then $h^* \in \mathcal{V}^0_t$. If we only care about removing edges of length $r = 2\nu + \varepsilon$ or more, so denote by $Q_r \subset Q$ the set of edges in Q with length at least r. We define the testing index to quantify informativeness of a test:

Definition 10 (Testing index)**.** *Let $\mathcal{P}$ be a class of joint probability distributions on $\mathcal{X} \times \mathcal{Y}$. A hypothesis class $\mathcal{H}$ is $(\rho, r, \Delta)_m$*-testable using m samples with respect to $\mathcal{P}$ *if for all finite edge sets $Q \subset \binom{H}{2}$, there exists a m-step sampling scheme S_m and a test T such that the test outcome ρ-splits Q_r with advantage Δ. That is, there exists T such that for all $P_{XY} \in \mathcal{P}$, the test outcome ρ-splits Q_r and*

$$\Pr_{T \sim \mathbb{P}^{(\mathsf{S}_m, P_{XY})}} [T(h^*) = 0] > \frac{1}{2} + \Delta.$$

Suppose that $\mathcal{H}$ is finite and $(\rho, r, \Delta)_m$-testable. Then, for any version space $\mathcal{V}$, there is a test T that can reduce the number of edges in $\mathcal{V}$ by $(1-\rho)$-fraction. In particular, suppose we want to compute a sequence of tests such that each edge of length r or more in $\mathcal{H}$ is split at least once. If we only query tests that ρ-splits the current version space, then at time s, we would have a version space $\mathcal{V}$ such that:

$$\begin{aligned} \#\text{ edges in } \mathcal{V} \text{ of length at least } r &\leq (1-\rho)^s \left| \binom{\mathcal{H}}{2} \right| \\ &\leq e^{-\rho s} \cdot |\mathcal{H}|^2. \end{aligned}$$

It follows that we can generate a sequence of no more than $\frac{2}{\rho} \log |\mathcal{H}|$ tests such that every edge is split by at least one test outcome. And if we wish to generate a sequence of tests such that each edge is split k times, then, we'd just repeat this procedure k times, using at most $\frac{2k}{\rho} \log |\mathcal{H}|$ tests. We claim that if k is sufficiently large, this sequence of tests is sufficient to reduce the diameter of the search space to at most r.

Before going on to prove this, though, let's understand the mechanics of these tests in the realizable setting, using a collection of the previously defined tests T_x, where $T_x(h)$ indicates whether $h(x)$ is wrong or not. From the realizability assumption, each test outcome is simply the map:

$$t_{(x, h^*(x))}(h) := \mathbf{1}[h(x) \neq h^*(x)].$$

Suppose the hypothesis class satisfies the following property: for all finite edge sets $Q \subset \binom{\mathcal{H}}{2}$, there exists $x \in \mathcal{X}$ such that T_x ρ-splits Q_r—regardless of how h^* labels x, upon obtaining its label, the resulting version space will have no more than $(1-\rho)$-fraction of the original edges. From the realizability assumption, T_x has advantage $\frac{1}{2}$, in that $T_x(h^*) = 0$ with probability 1. It then follows that $\mathcal{H}$ is $(\rho, r, \frac{1}{2})$-testable, and after performing no more than $\frac{2}{\rho} \log |\mathcal{H}|$ tests, we can remove all hypotheses connected to h^* by an edge with length at least r. We're thus left with a version space of diameter less than r, i.e., r-learning $\mathcal{H}$. And as each of the tests costs a single call to the LABEL oracle, the label complexity in this case is also $O\left(\frac{1}{\rho} \log |\mathcal{H}|\right)$.

But suppose we're no longer in the realizable setting, and the test outcomes merely have some positive advantage $0 < \Delta < \frac{1}{2}$. We could perform k tests for each edge in order to boost our advantage. Consider then the following observations:

1. For each edge $\{h^*, h\}$ of length at least r containing h^*, we perform k tests $T_1, \ldots, T_k$ such that:

   $$\Pr_{T_i} [T_i(h^*) = 1] < \frac{1}{2} - \Delta,$$

   which implies the following:

   $$\mathop{\mathbb{E}}_{T_1, \ldots, T_k} \left[\frac{1}{k} \sum_{i=1}^{k} T_i(h^*) \right] \leq \frac{1}{2} - \Delta.$$

2. For any edge $\{h, h'\}$ of length at least r, and their corresponding k tests $T_1, \ldots, T_k$, we have that:

   $$\frac{1}{k} \sum_{i=1}^{k} T_i(h) + T_i(h') \geq 1,$$

   since the tests are guaranteed to split the edge, and an edge $\{h, h'\}$ is split by T if and only if $T(h) + T(h') \geq 1$. It follows that one of h or h' must have empirical average of at least $\frac{1}{2}$.

These facts suggest that for each edge $\{h, h'\}$, we need to estimate $\frac{1}{k}\sum_{i=1}^{k} T_i(h)$ up to within $\Delta/2$ of its expected value. Then, for every edge $\{h^*, h\}$ of length at least r, we can confidently throw out the hypothesis h, whose empirical estimate is at least $\frac{1}{2}$. And for any edge $\{h, h'\}$ that do not contain h^*, it matters not which hypothesis we throw out; we'll apply the same rule from the $\{h^*, h\}$, removing any hypotheses that fail at least half of the tests. With high probability, we won't throw out h^*, while any edge of length at least r is deterministically removed; we're left with a version space $\mathcal{V}$ satisfying $h^* \in \mathcal{V} \subset B(h^*, r)$.

Algorithm *Test-based splitting*
($*$ diameter-based learning $*$)
Input: $\mathcal{H}$ hypothesis class, $r > 0$ an accuracy parameter, k the number of repetitions
Output: a version space $\mathcal{V}$
Initialize: $S_{h,h'} \leftarrow \varnothing$, for $(h, h') \in \mathcal{H} \times \mathcal{H}$
1. **for** $i = 1, \ldots, k$
2. **do** set preliminary version space $\mathcal{V} \leftarrow \mathcal{H}$
3. **while** $\operatorname{diam}(\mathcal{V}) \geq r$
4. **do** select and perform T a ρ-splitting test with advantage Δ on $\binom{\mathcal{V}}{2}_{\varepsilon}$
5. save the computation $S_{h,h'} \leftarrow S_{h,h'} \cup \{T(h)\}$, for $h, h' \in \mathcal{V}$
6. update the version space $\mathcal{V} \leftarrow \mathcal{V}_T^0$
7. set the final version space $\mathcal{V} \leftarrow \varnothing$
8. **for** $S_{h,h'}$ where $h \in \mathcal{H}$ and $d(h, h') \geq \varepsilon$
9. **do** compute $\mu_{h,h'}$ the average of values in $S_{h,h'}$
10. **if** $\mu_{h,h'} < \frac{1}{2}$ for all $h' \in \mathcal{H}$
11. **then** $\mathcal{V} \leftarrow \mathcal{V} \cup \{h\}$
12. **return** $\mathcal{V}$

To determine what to set k, we can apply the same analysis as before. We need to ensure that at most $|\mathcal{H}|$ estimates are accurate up to within $\pm\Delta/2$ with high probability: namely, the estimators $\frac{1}{k}\sum_{i=1}^{k} T_i(h^*)$ for each collection of k tests T^k associated with an edge $\{h^*, h\}$. It follows that we need $k = O\left(\frac{1}{\Delta^2}\log\frac{|\mathcal{H}|}{\delta}\right)$. Recalling that each test requires m calls to the LABEL oracle, we deduce the following:

Proposition 11 (Test-based splitting label complexity)**.** *Let $\mathcal{H}$ be $(\rho, r, \Delta)_m$-testable with respect to a class of joint probability distributions $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$. For any $P_{XY} \in \mathcal{P}$, the test-based splitting algorithm with input parameter $k = \Omega\left(\frac{1}{\Delta^2}\log\frac{|\mathcal{H}|}{\delta}\right)$ as above returns a subset $\mathcal{V} \subset \mathcal{H}$ such that $h^* \in \mathcal{V} \subset B(h^*, r)$ with probability $1 - \delta$. Additionally, it has label complexity at most:*

$$n = O\left(\frac{m}{\rho\Delta^2}\log|\mathcal{H}| \cdot \log\frac{|\mathcal{H}|}{\delta}\right).$$

Let's now extend our prior analysis of the collection of tests T_x from the realizable setting to an agnostic setting. Fixing a marginal distribution P_X on $\mathcal{X}$ (thus a pseudometric on $\mathcal{H}$), we can characterize the amenability of a hypothesis class to tests of the form T_x through the following *splitting index*.

Definition 12 (Splitting index)**.** *A hypothesis class $\mathcal{H}$ is* (ρ, r, τ)-splittable *if for all finite edge sets $Q \subset \binom{\mathcal{H}}{2}$, the probability mass of instances $x \in \mathcal{X}$ that correspond to ρ-splitting tests T_x on Q_r is at least τ:*

$$\Pr_{x \sim P_X}\left[T_x \text{ } \rho\text{-splits } Q_r\right] \geq \tau.$$

We call ρ the splitting index *of $\mathcal{H}$.*

Having fixed the pseudometric on $\mathcal{H}$, the splitting index becomes a purely combinatorial property of $\mathcal{H}$, where T_x ρ-splits an edge set Q if:

$$\max\left\{\left|Q \cap \binom{\mathcal{H}_x^+}{2}\right|, \left|Q \cap \binom{\mathcal{H}_x^-}{2}\right|\right\} \leq (1-\rho)|Q|,$$

where $\mathcal{H}_x^+$ is the collection of hypotheses that label x positively, and $\mathcal{H}_x^-$ is defined similarly. The splitting index quantifies the availability τ of instances $x \in \mathcal{X}$ that generate ρ-splitting tests T_x. In the following, as we're working in the agnostic setting, the tests T_x no longer have perfect advantage, as the LABEL oracle may label individual instances x arbitrarily badly, with respect to h^*. But we'll see that if $\nu = R(h^*)$ is not too large (i.e. the LABEL oracle provides information that is overall highly correlated to h^*) while τ is large (i.e. what constitutes useful information is not localized to a small number of noisy tests), then we may still see improvements to the label complexity. We claim the following:

Lemma 13. *Let P_X be a fixed marginal distribution over $\mathcal{X}$. Let $\mathcal{H}$ be a hypothesis class that is (ρ, r, τ)-splittable. Let $0 < \Delta \leq \frac{1}{2}$ and $0 \leq \nu < \left(\frac{1}{2} - \Delta\right)\tau$. Then, $\mathcal{H}$ is $(\rho, r, \Delta)_1$-testable over the following class of distributions:*

$$\mathcal{P} = \left\{P_X \otimes P_{Y|X} : \min_{h \in \mathcal{H}} R(P_{XY}, h) \leq \nu\right\}.$$

This lemma states that if we wish to be able to construct ρ-splitting tests of the form T_x that have positive advantage Δ for a hypothesis class with a fixed geometry, we'll be able to tolerate up to a certain amount of noise ν. In the worst case scenario, all the noise is localized to precisely the instances $x \in \mathcal{X}$ that correspond to ρ-splitting tests; the lemma follows from bounding those values of ν for which the noise is still tolerable. Applying Proposition 11, we obtain the immediate corollary:

Corollary 14. *Let $\mathcal{H}$ and $\mathcal{P}$ as before, so that $\mathcal{H}$ is (ρ, r, τ)-splittable and $\nu = R(h^*) < \tau/2$. There exists an active learning algorithm that (r, δ)-learns $\mathcal{H}$ over $\mathcal{P}$ with label complexity:*

$$n = O\left(\frac{\tau^2}{\rho(\tau - 2\nu)^2} \log|\mathcal{H}| \cdot \left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)\right).$$

For example, this shows that in the setting of low noise $\nu = o(\tau)$, we can learn, having a constant failure probability, with label complexities at most:

$$n_{\text{ACTIVE}} = O\left(\frac{1}{\rho}\log|\mathcal{H}|^2\right) \quad \text{vs.} \quad n_{\text{ERM}} = O\left(\frac{1}{\varepsilon}\log|\mathcal{H}|\right).$$

If it is the case that $\rho = \omega(\varepsilon)$ is not too small, then this improvement in label complexity can be significant with respect to the corresponding $O\left(\frac{1}{\varepsilon}\log|\mathcal{H}|\right)$ in the passive setting. Indeed, we do have the following lower bound, $\rho = \Omega(\varepsilon)$, which we prove in Appendix B:

Lemma 15 (Lower bound on ρ, [Das06])**.** *Let $0 < \alpha, \varepsilon < 1$. Let $\mathcal{V} \subset \mathcal{H}$. Then, $\mathcal{V}$ is $\big((1-\alpha)\varepsilon, \varepsilon, \alpha\varepsilon\big)$-splittable.*

Notice that the regime of Corollary 14 is essentially limited to $r > 2\nu$. In particular, whenever a hypothesis class is sufficiently expressive, the following upper bound on τ holds: $\tau \leq r$. To see this, consider any pair of hypotheses h, h' such that $d(h, h') = r$. In other words, they disagree on an r-fraction of $\mathcal{X}$. The size of any ρ-splitters of the singleton consisting of just the edge $\{h, h'\}$ has mass at most r. And since we require $\nu < \tau/2$, we must also have $r > 2\nu$.

It is for this reason that we divided the problem of (ε, δ)-learning $\mathcal{H}$ into two regimes: (i) the parametric regime where the structure of $\mathcal{H}$ overcomes the noisiness of the data and (ii) the non-parametric regime where the noise dominates any assumed structure. In the first regime, we aim to efficiently search for a version space $\mathcal{V} \subset B(h^*, r)$ where $r = 2\nu + \varepsilon$. In the second regime, we estimate the risk functional in order to (ε, δ)-learn $\mathcal{V}$. Returning to the original problem of (ε, δ)-learning $\mathcal{H}$, we obtain the following label complexity:

Theorem 16 (Splitting index label complexity)**.** *Let $\varepsilon, \delta > 0$ be fixed. Let $\mathcal{P}$ be a class of distributions for which the model error of $\mathcal{H}$ is at most ν:*

$$\max_{P_{XY}} \min_{h \in \mathcal{H}} R(P_{XY}, h) \leq \nu.$$

Let $\mathcal{H}$ have disagreement coefficient $\theta := \theta_{2\nu+\varepsilon}(\mathcal{H}, P_X, h^*)$ and splitting index $(\rho, 2\nu + \varepsilon, \tau)$. Then, there is an algorithm that (ε, δ)-learns $\mathcal{H}$ with label complexity at most:

$$n = O\left(\frac{1}{\rho}\frac{\tau^2}{(\tau - 2\nu)^2}\log|\mathcal{H}| + \theta^2\left(1 + \frac{\nu^2}{\varepsilon^2}\right)\right) \cdot \log\frac{|\mathcal{H}|}{\delta}.$$

3.3 Prelude to lower bounds: examples and counterexamples

3.3.1 Linear threshold classifiers

Perhaps the most cannonical example in active learning is the linear threshold classifier. Let $\mathcal{X} = [0, 1] \subset \mathbb{R}$ be the unit interval with uniform marginal distribution P_X. Define the hypothesis class $\mathcal{H}$ made up of step functions $h_\alpha(x) := \mathbf{1}\{x > \alpha\}$, for $\alpha \in [0, 1]$. It follows that $d(h_\alpha, h_{\alpha'}) = |\alpha - \alpha'|$, so that ε-learning h_{α^*} entails estimating α^* up to within $\pm\varepsilon$. For simplicity, let's assume we're in the noiseless, realizable setting, where we obtain data from h_{α^*} directly.

A passive learner must therefore sample enough data points $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ so that the x_i's forms a 2ε-cover for $[0, 1]$; otherwise, in the worst-case scenario, the threshold α^* falls into a range $[x_i, x_j]$ that is more than 2ε in diameter—there is no way to choose $\widehat{\alpha}$ from this range to guarantee that $d(\alpha^*, \widehat{\alpha}) < \varepsilon$. We previously saw that that the passive learner could learn this class of functions using $O\left(\frac{1}{\varepsilon}\right)$ queries. And so, the sample complexity of the passive learner here is:

$$n_{\text{PASSIVE}} = \Theta\left(\frac{1}{\varepsilon}\right).$$

An active learner, on the other hand, could perform a binary search for the threshold. The learner could cover $\mathcal{X}$ with an 2ε-net $\mathcal{N}$—every $h_\alpha \in \mathcal{H}$ is within ε of exactly one h_n for $n \in \mathcal{N}$. It takes at least $\log\frac{1}{2\varepsilon}$ bits to identify any single $h_n \in \mathcal{N}$, so we need at least that many queries. And in fact, a binary search achieves this lower bound; the label complexity of active learning $\mathcal{H}$ is:

$$n_{\text{ACTIVE}} = \Theta\left(\log\frac{1}{\varepsilon}\right).$$

This cannonical example of active learning shows that when data can be collected sequentially, the amount of data required could decrease exponentially from the passive learning setting. The possibility for improvement here can be explained both through the disagreement coefficient and the splitting index.

Lemma 17. *Let $\mathcal{H}$ be the class of linear threshold classifiers on $\mathcal{X} = [0, 1]$ with marginal distribution P_X. Then, the disagreement coefficient is constant $\theta_s(\mathcal{H}, P_X, h^*) \leq 2$. Furthermore, it is $\left(\frac{1}{2}, \varepsilon, \varepsilon\right)$-splittable.*

However, not all hypothesis classes have geometry as amenable to learning as linear threshold classifiers. Next, we consider a classes of hypotheses where a brute-force search is required in order to ε-learn.

3.3.2 Interval functions

Consider the collection of interval functions on unit interval,

$$h_{\alpha,\beta}(x) := \mathbf{1}\{\alpha < x < \beta\},$$

for $0 \leq \alpha < \beta \leq 1$. For simplicity, assume that P_X has the uniform distribution. Learning this class of functions is at least as hard as learning the subset:

$$\mathcal{H}' = \left\{ h_{k\varepsilon,(k+1)\varepsilon} : 0 \leq k < \frac{1}{\varepsilon} \right\},$$

the class of hypotheses with pairwise disjoint support, with interval lengths ε. Because this subset is a 2ε-packing of $\mathcal{H}$, to ε-learn this subclass means to exactly recover $h_{k^*\varepsilon,(k^*+1)\varepsilon}$. For any querying strategy, in the worst case scenario, we will need exactly $\frac{1}{\varepsilon} - 1$ queries before encountering the interval where the true hypothesis is supported. On the other hand, the VC dimension of $\mathcal{H}$ is 2, and so in the noiseless setting, from Theorem 39, we see that the label complexity for this class is:

$$n_{\text{PASSIVE}} = \Theta\left(\frac{1}{\varepsilon}\right) = n_{\text{ACTIVE}}.$$

If we're not in the worst-case scenario, then perhaps $h^* = h_{\alpha,\beta}$ where $\gamma = \beta - \alpha$ is much larger than ε. Then, notice that once we've found the first point $x \in [0,1]$ that is labeled 1, we could perform a binary search for the two ends of the interval. It follows that an active learner could would require at most:

$$n_{\text{ACTIVE}} = O\left(\frac{1}{\gamma} + \log\frac{1}{\varepsilon}\right),$$

which is asymptotically $\log\frac{1}{\varepsilon}$ instead of $\frac{1}{\varepsilon}$. This is not unlike our earlier bookstore example, where initially, there is nothing better to do but a brute-force search for the right section, but once we found the right section, we could perform a binary search on the book author.

This is also reflected in the disagreement coefficient and splitting index:

Lemma 18. *Let $\mathcal{H}$ be the class of interval functions on $\mathcal{X} = [0,1]$, with marginal distribution P_X. Let $h^* = h_{\alpha,\beta}$, where $\gamma = \beta - \alpha$. The disagreement coefficient is:*

$$\theta_s(\mathcal{H}, P_X, h_{\alpha,\beta}) = \begin{cases} \frac{1}{s} & s > \gamma \\ 4 & s \leq \gamma. \end{cases}$$

For any $0 < s < 1$, $\mathcal{H}$ is not $(\rho, \varepsilon, \tau)$-splittable for any $\rho > 2\varepsilon$.

While the geometry of a hypothesis class plays a great deal in its learnability, the amount of noise or the degree to which the model was misspecified is another integral aspect to how difficult it is to learn. Now, we'll return to an example we briefly mentioned at the start of this section: the binary hypothesis testing problem.

3.3.3 Binary hypothesis testing

Let's consider the problem of (ε, δ)-learning the hypothesis class $\mathcal{H} = \{\mathbf{0}, \mathbf{1}\}$ over some instance space $\mathcal{X}$. Equivalently, we can think of this as a binary hypothesis testing problem: for example, suppose we want to determine whether a coin is biased toward *heads* or *tails* (let's say these are 1 and 0, respectively). We let the outcome of a coin flip be the random variable $X \sim \text{Ber}(p)$, where p is the probability of landing on heads.

In the context of ε-learning, we only care to know if the coin biases one outcome over the other at least an ε-fraction of the time. That is, when $\left|p - \frac{1}{2}\right| > \frac{1}{2}\varepsilon$. It follows from our early discussion on estimation that we need only to estimate p up to within $\frac{1}{2}\varepsilon$. To do this with constant failure probability, depending on the margin between p and $\frac{1}{2}$, we potentially need up to n flips, with:

$$n = O\left(\frac{1}{\varepsilon^2}\right).$$

Indeed, if the margin is $\frac{1}{2}$ (so the coin always returns the same outcome), then a single flip is sufficient to determine which way the coin is biased (i.e. a single query to LABEL is enough to ε-learn $\mathcal{H}$). But if the margin approaches $\varepsilon/2$, then in fact, the previous $\frac{1}{\varepsilon^2}$ label complexity is tight. We'll show this in two ways. The first is direct, using Slud's lemma. The second will preview a more general information-theoretic technique to prove lower bounds for active learning. In contrast to Hoeffding's inequality, Slud's lemma lower bounds the following probability of failure:

Lemma 19 (Slud's lemma)**.** *Let* $X_1, \dots, X_n \sim \mathrm{Ber}\left(\frac{1}{2} - \frac{1}{2}\varepsilon\right)$*. Then:*

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i > \frac{1}{2}\right] \geq \frac{1}{2}\left(1 - \sqrt{1 - \exp\left(\frac{-n\varepsilon^2}{1-\varepsilon^2}\right)}\right).$$

It follows that if $\varepsilon \ll 1$, then in order to obtain a constant probability of failure when determining which way the coin is biased, we need at least n samples, with:

$$n = \Omega\left(\frac{1}{\varepsilon^2}\right).$$

This shows that the presence of noise can make learning difficult, even over a hypothesis class as structured as $\mathcal{H} = \{\mathbf{0}, \mathbf{1}\}$. Put another way, we have two distributions—$\mathrm{Ber}\left(\frac{1}{2} - \frac{1}{2}\varepsilon\right)$ and $\mathrm{Ber}\left(\frac{1}{2} + \frac{1}{2}\varepsilon\right)$—that are hard to distinguish in an information-theoretic sense, requiring us to estimate the Bernoulli parameter up to $\varepsilon/2$. Yet, they are still mapped to different hypotheses when they are learned correctly. To elaborate on the information-theoretic approach, recall the *total variation* of two probability measures on a common space:

Definition 20 (Total variation distance)**.** *Let* P_0 *and* P_1 *be two probability measures on a sample space* Ω*. Their* total variation distance *is:*

$$\mathrm{TV}(P_0, P_1) := \sup_{A \in \mathcal{F}} |P_0(A) - P_1(A)|,$$

where the supremum is taken over the set of all events $\mathcal{F}$*.*

The total variation distance characterizes what it means for two distributions to be 'hard to distinguish'. Consider the following Bayesian problem: nature choose one of two distributions P_0 and P_1 over the same outcome space Ω uniformly at random, and presents a sample X drawn from that distribution. Can we perform a test $\Psi : \Omega \to \{0, 1\}$ so that just from looking at the outcome X, we can recover which distribution it was drawn from reasonably well?

Given any test, we can define its *probability of error* as the probability that it fails to recover the true $\theta \in \{0, 1\}$ from which the sample $X \sim P_\theta$ was drawn. Then the probability of error is:

$$\frac{1}{2}P_0(\Psi(X) \neq 0) + \frac{1}{2}P_1(\Psi(X) \neq 1).$$

The total variation distance characterizes the indistinguishability of P_0 and P_1 in the following way:

Lemma 21 (Variational formulation [Duc14])**.** *Let* P_0 *and* P_1 *be two distributions on* Ω*. The total variation distance satisfies:*

$$\mathrm{TV}(P_0, P_1) = 1 - \inf_{\Psi}\left\{P_0(\Psi(X) \neq 0) + P_1(\Psi(X) \neq 1)\right\},$$

where the infimum is taken over all measurable functions $\Psi : \Omega \to \{0, 1\}$*.*

In particular, by rearranging, we see that the error rate of any test is lower bounded:

$$\inf_{\Psi} \Pr_{\theta, X}\left[\Psi(X) \neq \theta\right] = \frac{1}{2} - \frac{1}{2}\mathrm{TV}(P_0, P_1).$$

Let $P_0 = \text{Ber}\left(\frac{1}{2} - \frac{1}{2}\varepsilon\right)$ and $P_1 = \text{Ber}\left(\frac{1}{2} + \frac{1}{2}\varepsilon\right)$. Suppose that we have an ε-learner for $\mathcal{H}$ that has label complexity n. That is, given n i.i.d. samples from the data distribution, this learner is able to recover which distribution P_0 or P_1 it came from; it must be able to distinguish between the product distributions, $P_0^{\otimes n}$ and $P_1^{\otimes n}$. It follows that the ε-learner can be converted into a hypothesis tester for $P_0^{\otimes n}$ and $P_1^{\otimes n}$.

Conversely, we can convert this lower bound on the probability of error to lower bounds on the label complexity of learning $\mathcal{H}$. The above bound shows that there does not exist any algorithm that can distinguish $P_0^{\otimes n}$ and $P_1^{\otimes n}$ with error rate δ less than:

$$\delta < \frac{1}{2} - \frac{1}{2}\text{TV}(P_0^{\otimes n}, P_1^{\otimes n}).$$

Thus, any (ε, δ)-learner for $\mathcal{H}$ must sample enough points n so that the induced product distributions become well-separated with respect to the total variation distance, $\text{TV}(P_0^{\otimes n}, P_1^{\otimes n}) > 1 - 2\delta$.

But, instead of computing $\text{TV}(P_0^{\otimes n}, P_1^{\otimes n})$ directly, let us recall another notion of 'distance' for probability measures—the KL-divergence:[4]

Definition 22 (KL-divergence)**.** *Let P and Q be two distributions defined on a finite space $\mathcal{X}$. The* Kullback-Leibler (KL) divergence*,* $\text{KL}(P\|Q)$*, is defined as:*

$$\text{KL}(P\|Q) := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)},$$

where we interpret $\log \frac{0}{0}$ *as* 0*.*

The total variation distance of two distributions is often difficult to bound directly; but, if we can bound their KL-divergence, we can upper bound the former via Pinsker's inequality:

Lemma 23 (Pinsker's inequality)**.** *Let P and Q be two probability distributions on $\mathcal{X}$. Then:*

$$\text{TV}(P, Q) \leq \sqrt{\frac{1}{2}\text{KL}(P\|Q)}.$$

In fact, the KL-divergence of a pair of product distributions has a particularly nice form:

Lemma 24 (KL-divergence of a product)**.** *Let $P_1, \ldots, P_n$ and $Q_1, \ldots, Q_n$ be distributions over $\mathcal{X}_1, \ldots, \mathcal{X}_n$. Let $\mathbb{P} := \bigotimes_{i=1}^n P_i$ and $\mathbb{Q} := \bigotimes_{i=1}^n Q_i$ be the product distributions over $\prod_{i=1}^n \mathcal{X}_i$. Their KL-divergence is:*

$$\text{KL}(\mathbb{P}\|\mathbb{Q}) = \sum_{i=1}^n \text{KL}(P_i\|Q_i).$$

It follows that the sample distribution for n independent draws has total variation at most:

$$\text{TV}(P_0^{\otimes n}, P_1^{\otimes n}) \leq \sqrt{\frac{n}{2}\text{KL}(P_0\|P_1)}.$$

The KL-divergence for two Bernoulli random variables can be bounded:

Lemma 25 (KL-divergence of Bernoulli variables, [CN08])**.** *Let $P = \text{Ber}\left(\frac{1}{2} - p\right)$ and $Q = \text{Ber}\left(\frac{1}{2} - q\right)$, where $|p|, |q| < \frac{1}{4}$. Then:*

$$\text{KL}(P\|Q) \leq 8(p - q)^2.$$

[4]The KL-divergence is not a proper distance because it is not symmetric; it does satisfy reflexivity, positivity, and triangle inequality, however.

Letting $p = -\varepsilon/2$ and $q = \varepsilon/2$, we see that an (ε, δ)-learner must have label complexity at least n, where n is sufficiently large so that it satisfies:

$$\delta > \frac{1}{2} - \sqrt{n\varepsilon^2}.$$

Once again, we see that to ensure a constant probability of failure, any (ε, δ)-learner has label complexity:

$$n = \Omega\left(\frac{1}{\varepsilon^2}\right).$$

3.3.4 Simple lower bound using VC dimension

As a simple extension of the lower bound we derived for the binary hypothesis testing scenario, we can consider any hypothesis class with VC dimension V. In particular, if $\mathcal{H}$ has VC dimension V, there exist a collection of points $x_1, \ldots, x_V \in \mathcal{X}$ that is shattered by $\mathcal{H}$. In the noiseless setting, a learner must query the labels for each of these V points. However, suppose that we're now in the setting with noise, where $R(h^*) = \nu$. Then for a learner to determine how h^* labels any single point x_i, it must solve the binary hypothesis testing problem for x_i. The following lower bounds shows that in some cases, any ε-learner must successfully perform a binary hypothesis testing for a constant fraction of those V points:

Proposition 26 (Theorem 12, [BDL08])**.** *Fix $0 \leq \nu \leq \frac{1}{2}$. Let $\mathcal{H}$ be a hypothesis class with VC dimension V. There exists a class of distributions $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ such that $R(h^*) = \nu$ and any algorithm that (ε, δ)-learns $\mathcal{H}$ over $\mathcal{P}$ must have label complexity at least:*

$$n = \Omega\left(V \cdot \frac{\nu^2}{\varepsilon^2}\right).$$

4 Lower bounds for active learning

We've seen now that we can divide the learning process into two regimes: one where noise dominates the structural assumptions made by the hypothesis class, and one where the underlying data distribution satisfies those assumptions relatively well. And so, we'll also have two types of lower bounds on the label complexity of learning: (i) an information-theoretic lower bound based on how difficult it is to distinguish between two similarly noisy distributions that lead to very different risk minimizers, and (ii) a geometric lower bound based on the number of dimensions/degrees of freedom on which two hypotheses in $\mathcal{H}$ can differ.

Back at the bookstore, the first type of bound corresponds to the setting where books are disorderly and misplaced; perhaps the high-level sections remain relatively consistent, while within each section, books have little discernible order. The second type of bound corresponds to the setting where each shelf corresponds to a different section, and one just has to come across the right shelf before beginning learning through a more intelligent or binary-search-like approach.

4.1 Information-theoretic lower bounds

To prove lower bounds from an information-theoretic framework, we'll make use of standard non-parametric techniques, described for example in [Tsy09] and [Duc14]. We already saw a simple example of this when we lower bounded the label complexity of binary hypothesis testing. Here, we'll describe a more general technique, which also reduces a particular hard hypothesis testing problem into a learning problem: given an (ε, δ)-learning algorithm $\mathcal{A}(\mathcal{P}, \mathcal{H}, \varepsilon, \delta)$ with label complexity $m_{\mathcal{A}}$, we convert the algorithm into one that solves the hypothesis testing problem. Because we can then show that there exists a learning problem whose solution implies the solution to a hard hypothesis testing problem, any lower bound applying to the testing

problem then applies to the learning problem—we obtain a *minimax lower bound* for the (ε, δ)-learning problem:

$$\min_{\mathcal{A}} \max_{\mathcal{P},\mathcal{H}} m_{\mathcal{A}(\mathcal{P},\mathcal{H},\varepsilon,\delta)}.$$

That is, how many labels does even the best possible learning algorithm need in the worst-case scenario? Before we make the connection to learning, let's describe in more detail one can more generally obtain information-theoretic lower bound.

To describe the hypotheis testing problem, let us construct a universe $\mathcal{U}$ where the underlying data is generated by one out of $N+1$ possible distributions:

$$\mathsf{P} = \{P_0, \ldots, P_N\}.$$

The hypothesis testing problem entails determining, with high probability, which one of the $N+1$ distributions was in fact realized. In other words, let $\theta \in \Theta \equiv \{0, \ldots, N\}$ parametrize our possible world. Our goal is to recover θ based on access to the samples from the distribution P_θ. To do this, we can define a test $\Psi : \mathcal{U} \to \Theta$, whereupon observing the outcome of a random draw $U \sim P$, the test predicts which one of the underlying distribution it came from, $\widehat{\theta} := \Psi(U)$. We define the *minimax probability of error* p_e as:

$$p_e := \min_{\Psi} \max_{\theta \in \{0,\ldots,N\}} P_\theta(\Psi(U) \neq \theta).$$

This value is the error rate of the best test in the worst-case scenario. Intuitively, if the distributions in P are quite similar, then the minimax probability of error will be high, and it will be impossible to distinguish them using a single sample with a great amount of certainty. In order to test between these hypotheses $\theta \in \Theta$ with greater success, we could attempt to transform the distributions from which we obtain data. For example, in our binary hypothesis testing discussion, we drew multiple i.i.d. samples repeatedly from P, so that instead of having to distinguish $P_0, \ldots, P_N$, we could distinguish the product distributions $P_0^{\otimes m}, \ldots, P_N^{\otimes m}$. As the number of samples m grows, the product distributions also grow apart; at some point, it becomes possible to recover θ with high confidence.

4.1.1 Reduction from hypothesis testing

Imagine that we had an algorithm $\mathcal{A}$ that could (ε, δ)-learn $\mathcal{H}$ over a class of distributions $\mathcal{P}$ with a fixed marginal distribution P_X using m samples. Let S_m be its m-step learning strategy, so that the learner samples its training data set from $\mathbb{P}^{(\mathsf{S}_m, P_{Y|X})}$ over $(\mathcal{X} \times \mathcal{Y})^m$; that is, the training dataset $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ is sampled with probability:

$$\mathbb{P}^{(\mathsf{S}_m, P_{Y|X})}((x_1, y_1), \ldots, (x_m, y_m)) = \prod_{i=1}^{m} \Pi_{X_i|X^{i-1},Y^{i-1}}(x_i|x^{i-1}, y^{i-1}) P_{Y|X}(y_i|x_i).$$

Depending on which sequence of training data in $(\mathcal{X} \times \mathcal{Y})^m$ is sampled, the learner returns a different hypothesis $h \in \mathcal{H}$.

We can convert $\mathcal{A}$ into an algorithm for the following hypothesis testing problem: consider $N+1$ possible worlds $\theta \in \Theta := \{0, \ldots, N\}$ corresponding to $N+1$ different conditional distributions $P_{Y|X,0}, \ldots, P_{Y|X,N}$. Suppose that it is the case that any learner $\mathcal{A}$ who ε-learns $\mathcal{H}$ correctly with respect to these data distributions will necessarily learn distinct hypotheses for distinct distributions. In other words, if a hypothesis h is ε-close to a risk minimizer with respect to one of the distributions $P_X \otimes P_{Y|X,\theta}$, then it is not ε-close to a risk minimizer for any other distributions $P_X \otimes P_{Y|X,\theta'}$.

And so, a learner who can (ε, δ)-learn in this setting must be able to distinguish between the different induced sampling distributions $\mathbb{P}^{(\mathsf{S}_m, P_{Y|X,\theta})}$, for $\theta \in \Theta$ with probability of error at most δ. Therefore, if we

can bound how quickly the sampling distributions diverge from each other as the number of samples m increases, then we can lower bound the number of samples required by any (ε, δ)-learner in this setting.

Our method of attack to obtain label complexity lower bounds for an (ε, δ)-learner over a hypothesis class $\mathcal{H}$ with respect to a class of distribution $\mathcal{P}$ will be to find a collection of distributions in $\mathcal{P}$ that are information-theoretically hard to distinguish, yet still map to disjoint sets of hypotheses in $\mathcal{H}$ by an ε-learner. Then, any hypothesis testing lower bound can be translated into an active learning lower bound. As a first simplification, note that:

$$p_e = \min_{\Psi} \max_{\theta} P_\theta(\Psi \neq \theta) \geq \min_{\Psi} \frac{1}{N+1} \sum_{\theta=0}^{N} P_\theta(\Psi \neq \theta).$$

Instead of lower bounding p_e directly, we'll lower bound the latter quantity. One interpretation of this quantity is the *average probability of error* in a Bayesian setting: the parameter θ is chosen uniformly at random from $\{0, \ldots, N\}$, and our goal is recover θ from the distribution P_θ well on an average case basis.

In fact, our analysis of binary hypothesis testing did just this: there $N = 1$, and our lower bound for the average probability of error also translates to the minimax probability of error, which we derived from the following sequence of inequalities:

$$\begin{aligned} p_e &\geq \inf_{\Psi} \frac{1}{2} P_0(\Psi \neq 0) + \frac{1}{2} P_1(\Psi \neq 1) \\ &= \frac{1}{2} - \frac{1}{2} \mathrm{TV}(P_0, P_1) \\ &\geq \frac{1}{2} - \frac{1}{2} \sqrt{\frac{1}{2} \mathrm{KL}(P_1 \| P_0)}. \end{aligned}$$

And so, the following holds generally:

Proposition 27 (Le Cam's method)**.** *Let P_0 and P_1 be probability measures on $\mathcal{X}$ such that $P_1 \ll P_0$. If the following holds:*

$$\frac{1}{2} \mathrm{KL}\left(P_1 \| P_0\right) < \alpha,$$

then we have:

$$p_e \geq \frac{1}{2} - \frac{1}{2} \sqrt{\alpha}.$$

This suggest the following approach to lower bounding the label complexity of an active learning algorithm $\mathcal{A}$ that (ε, δ)-learns a hypothesis class $\mathcal{H}$ over a class of distributions $\mathcal{P}$:

1. Let $P_0 \equiv P_X \otimes P_{Y|X,0}$ and $P_1 \equiv P_X \otimes P_{Y|X,1}$ be two data distributions over $\mathcal{X} \times \mathcal{Y}$ from $\mathcal{P}$ such that any hypothesis that no hypothesis $h \in \mathcal{H}$ can simultaneously be a solution to ε-learning both distributions. That is, $\mathcal{H}_0$ and $\mathcal{H}_1$ are disjoint, where $\mathcal{H}_\theta$ for $\theta \in \{0, 1\}$ are defined as:

$$\mathcal{H}_\theta := \left\{ h \in \mathcal{H} : R(P_\theta, h) < \min_{h' \in \mathcal{H}} R(P_\theta, h') + \varepsilon \right\}.$$

2. Upper bound the KL-divergence of the induced sample distributions $\mathbb{P}^{\mathsf{S}_m, P_{Y|X,0}}$ and $\mathbb{P}^{\mathsf{S}_m, P_{Y|X,1}}$, where S_m is the m-step learning strategy that $\mathcal{A}$ uses (i.e. show that there is a limit to how well-separated the sample distributions that the m-step learning strategy can induce):

$$\frac{1}{2} \mathrm{KL}\left(\mathbb{P}^{\mathsf{S}_m, P_{Y|X,1}} \middle\| \mathbb{P}^{\mathsf{S}_m, P_{Y|X,0}}\right) < \alpha_m.$$

3. The label complexity of (ε, δ)-learning $\mathcal{H}$ over $\mathcal{P}$ is then at least:

$$m_{\text{ACTIVE}} \geq \min\left\{m \in \mathbb{N} : \delta < \frac{1}{2} - \frac{1}{2}\sqrt{\alpha_m}\right\}.$$

Of course, different choices of P_X, $P_{Y|X,0}$ and $P_{Y|X,1}$ yields better or worse lower bounds on the label complexity. In fact, the KL-divergence of the sample distributions are related to their conditional distributions; we can upper bound the former with respect to the latter. The following lemma shows that if the divergence between $P_{Y|X=x,0}$ and $P_{Y|X=x,1}$ is large, then a learning strategy that increases the expected number of queries to the point x will also increase the divergence between $\mathbb{P}^{\mathsf{S}_m, P_{Y|X,0}}$ and $\mathbb{P}^{\mathsf{S}_m, P_{Y|X,1}}$.

Lemma 28 (KL-divergence of sequentially sampled data, [RR11])**.** *Let S_m be a m-step learning strategy and let $P_{Y|X,\theta'}$ and $P_{Y|X,\theta}$ be two conditional distributions. Suppose that they induce the sample distributions $\mathbb{P}^{\mathsf{S}_m, P_{Y|X,\theta'}}$ and $\mathbb{P}^{\mathsf{S}_m, P_{Y|X,\theta}}$. Then:*

$$\mathrm{KL}\left(\mathbb{P}^{\mathsf{S}_m, P_{Y|X,\theta'}} \middle\| \mathbb{P}^{\mathsf{S}_m, P_{Y|X,\theta}}\right) = \sum_{x \in \mathcal{X}} N_{\theta'}(x) \cdot \mathrm{KL}\left(P_{Y|X=x,\theta'} \middle\| P_{Y|X=x,\theta}\right),$$

where $N_{\theta'}(x) = \mathbb{E}_{\mathbb{P}^{\mathsf{S}_m, P_{Y|X,\theta'}}}\left[\#\{t : x_t = x\}\right]$ is the expected number of times x will be queried when given responses from $P_{Y|X,\theta'}$.

Applying Hölder's inequality, we immediately obtain the following, which we may apply toward compute a label complexity lower bound using step 3 from above:

Corollary 29 (Upper bound on KL divergence, [WS16])**.** *Let S_m, $P_{Y|X,\theta'}$ and $P_{Y|X,\theta}$ as before. Then, the following holds:*

$$\mathrm{KL}\left(\mathbb{P}^{\mathsf{S}_m, P_{Y|X,\theta'}} \middle\| \mathbb{P}^{\mathsf{S}_m, P_{Y|X,\theta}}\right) \leq m \cdot \sup_{x \in \mathcal{X}} \mathrm{KL}\left(P_{Y|X=x,1} \middle\| P_{Y|X=x,0}\right).$$

Given a certain hypothesis class and noise assumptions, there may be a limit to how close two distributions $P_{Y|X=x,0}$ and $P_{Y|X=x,1}$ can get for all $x \in \mathcal{X}$. Instead of reducing a binary hypothesis testing problem to a learning one, as we have just done here, the following allows us to produce lower bounds by reducing a multiple hypothesis testing problems. In the application of the technique outlined above, the only difference is the construction of $N+1$ distributions $P_0, \ldots, P_N$ that correspond to $N+1$ disjoint hypothesis class $\mathcal{H}_0, \ldots, \mathcal{H}_N$.

Proposition 30 (Proposition 2.3, [Tsy09])**.** *Let $P_0, \ldots, P_N$ be probability measures on $\mathcal{X}$ with $N \geq 2$ such that $P_i \ll P_0$. If the following holds:*

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{KL}(P_i \| P_0) \leq \alpha \log N,$$

with $0 < \alpha < \infty$, then:

$$p_e \geq \left[\frac{\sqrt{N}}{1+\sqrt{N}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log N}}\right)\right]$$

The information-theoretic lower bound technique we've discussed so far is particularly useful to analyze settings with significant amounts of noise or specific types of noise. In the next section, we'll define a combinatorial quantity that provide lower bounds that arise from the geometry of the hypothesis class, capturing our earlier example of interval functions.

4.2 Geometric lower bounds

Recall back to the example hypothesis class of interval functions—in particular, we considered a subcollection of interval functions of interval lengths ε with disjoint support:

$$\mathcal{H}' = \left\{ h_{k\varepsilon,(k+1)\varepsilon} : 0 \le k \le \frac{1}{\varepsilon} \right\}.$$

In order to ε-learn this subclass of hypotheses, a learner must be able to rule out all hypotheses but the target hypothesis. On the other hand, every point potentially only rules out a single hypothesis. Thus, in the worst-case, a learner must use $\Omega\left(\frac{1}{\varepsilon}\right)$ queries to the LABEL oracle to brute-force search for the interval. This example motivates the following quantity, as defined by [HY15]:

Definition 31 (Star number)**.** *The* star number $\mathfrak{s}$ *of a hypothesis class $\mathcal{H}$ is the largest integer s where there exists $x_1, \dots, x_s \in \mathcal{X}$ and $h_0, h_1, \dots, h_s \in \mathcal{H}$ such that:*

$$\mathrm{DIS}(\{h_0, h_i\}) \cap \{x_1, \dots, x_s\} = \{x_i\}.$$

We can generalize difficulty of learning $\mathcal{H}'$ to any hypothesis class with a large star number. In particular, given a hypothesis class $\mathcal{H}$ with a star number $\mathfrak{s}$, we can construct a class of distributions where at least $\min\left\{\mathfrak{s}, \frac{1}{\varepsilon}\right\} \equiv \mathfrak{s} \wedge \frac{1}{\varepsilon}$ labels are required in the worst case. We'll give the construction in the realizable setting.

Let $k = \mathfrak{s} \wedge \frac{1}{\varepsilon}$. Define the marginal distribution P_X to be uniformly distributed on its support, consisting only of $x_1, \dots, x_k$. Thus, every x_i has mass at least ε. When we view $\mathcal{H}$ as a graph, every edge $\{h_0, h_i\}$ has length at least ε, and the only point that certainly splits that edge is x_i. It follows that in the worst case scenario, we have a label complexity of k. Combining this geometry-based lower bound with the information-theoretic lower bound of Proposition 26, we obtain:

Proposition 32 (Active learning lower bound, [HY15])**.** *Fix $0 \le \nu < \frac{1}{2}$ and $0 < \varepsilon < \frac{1}{24}(1 - 2\nu)$. If $\mathcal{H}$ is a hypothesis class with VC dimension V and star number $\mathfrak{s}$, there exists a class of distributions $\mathcal{P}$ where $\min_{h \in \mathcal{H}} R(P, h) \le \nu$ for all $P \in \mathcal{P}$, and if any active learner $\mathcal{A}$ that ε-learns $\mathcal{H}$ over $\mathcal{P}$, then it must have label complexity at least:*

$$n_{\mathcal{A}} = \Omega\left(V \cdot \frac{\nu^2}{\varepsilon^2} + \mathfrak{s} \wedge \frac{1}{\varepsilon} \right).$$

5 Epilogue

In this set of notes, we presented two approaches to active learning: through (i) estimating the relative risks of a hypothesis class and (ii) estimating the labels of highly informative instances. In both of these cases, we thought of learning as performing experiments—by designing a statistic that furthers the goal of ε-learning $\mathcal{H}$, before obtaining data to estimate the statistic.

We presented learning in this way in order to suggest the possibility that there are other statistics besides the ones we studied that may be applied to active learning and its analysis. The presentation went only as far as needed to frame active learning in a way such to give the reader knowledge in designing and analyzing new active learning algorithms. We'll briefly provide remarks on a few results that could have been tightened, in addition to a few directions for further research.

1. In most of the results, we simplified arguments by allowing $\mathcal{H}$ to be a finite hypothesis class. To extend results to infinite hypothesis classes, we could construct a sufficiently fine covering of $\mathcal{H}$. In particular, the size of an ε-covering is on the order of $\left(\frac{1}{\varepsilon}\right)^V$, where V is the VC dimension (see Theorem 44). Because we've approximated $\mathcal{H}$ with a cover $\mathcal{H}'$, the true risk of the new best-in-class $h^* \in \mathcal{H}'$ may

increase by $O(\varepsilon)$. Then, our results for finite hypothesis classes may be extended to infinite hypothesis classes by replacing $\log|\mathcal{H}|$ with $O\left(V \log \frac{1}{\varepsilon}\right)$ and ν with $O\left(\nu + \varepsilon\right)$.

2. We defined the testing index $(\rho, \varepsilon, \Delta)_m$ with respect to tests that deterministically split a ρ-fraction of edges. We can relax this by allowing them to probabilistically split edges, with failure rate at most δ_ρ. We can modify our testing algorithm to learn using tests that non-deterministically ρ-split edges by accounting for this additional failure probability.

3. This work does not attempt to achieve tight label complexity bounds. For example, Theorem 16 providing the label complexity upper bound for the splitting index algorithm in the agnostic setting can likely be tightened—it is loose when compared to the corresponding upper bound for the realizable setting demonstrated in [Das06].

5.1 Bibliographic remarks

This work was heavily based on a previous survey on active learning, [Das11]. It is also there that certain techniques of active learning were classified into mellow or aggressive approaches.

The A^2 algorithm is originally designed and analyzed in [BBL09], which extends an early active learning algorithm from [CAL94], operating within the realizable setting, to an agnostic setting. The label complexity analysis that we provide is given by [Han+14]. Recall that in A^2, we performed $\log \frac{1}{\varepsilon}$ rounds of independent estimations. We can, in fact, obtain a more efficient active learning algorithm if all the data we label are used to compute a running risk estimation. As we draw and label points to estimate the risk functional, at certain points, it will become clear that the risk of a particular hypothesis h is much larger than the risk of the best-in-class hypothesis; we can throw h out of our version space to shrink the disagreement region. Now, whenever a point x is not in the disagreement region, we can forgo querying its label, since all remaining hypotheses agree on it anyways—instead, we just infer its label and incorporate that into our running risk estimate. Regardless of the correctness of the inferred label, the relative risks of the hypotheses in the version space are preserved. The algorithm that sequentially pares down $\mathcal{H}$ in this way is designed in [DHM08], and it obtains an improved label complexity of:

$$n_{\text{DHM}} = \widetilde{O}\left(d\theta\left(\log^2 \frac{1}{\varepsilon} + \frac{\nu^2}{\varepsilon^2}\right)\right).$$

The testing index was inspired both by [Das06] and [RR11]. In particular, our definition of a learning strategy is borrowed from [RR11], and the testing index generalizes the splitting index in [Das06]. There, the analysis was performed in the realizable setting, and obtained a label complexity of:

$$n_{\text{SPLITTING}} = \widetilde{O}\left(\frac{1}{\rho} \log \frac{|\mathcal{H}|}{\delta}\right).$$

The presentation in the binary hypothesis testing drew from [Duc14], from which we obtained a lower bound of $\Omega\left(\nu^2/\varepsilon^2\right)$. See also [Kää06] and [BDL08] for more discussion on lower bounds with respect to the noise rate. The more general approach to lower bounding label complexity follows [CN08] and [RR11], which uses standard techniques as described in [Tsy09] and [Duc14]. For a good survey of the three main approaches, see also [Yu97].

Throughout this survey, we've imposed no conditions on the noise or model misspecification, except that:

$$\min_{h \in \mathcal{H}} R(h) = \nu.$$

But in certain situations, we might make some further assumptions about how the noise is distributed—for example, certain regions of $\mathcal{X}$ have noisier labels than others. Common examples of such assumptions include the Massart/bounded noise condition, Tsybakov noise condition, and Bernstein noise condition. As examples, [RR11] provides lower bounds on the label complexity under Massart noise condition and [WS16] for Tsybakov noise condition for a specific hypothesis class.

References

[Hau92] David Haussler. "Decision theoretic generalizations of the PAC model for neural net and other learning applications". In: *Information and Computation*. 1992, pp. 78–150.

[CAL94] David Cohn, Les Atlas, and Richard Ladner. "Improving generalization with active learning". In: *Machine learning* 15.2 (1994), pp. 201–221.

[Yu97] Bin Yu. "Assouad, fano, and le cam". In: *Festschrift for Lucien Le Cam*. Springer, 1997, pp. 423–435.

[BBL03] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. "Introduction to statistical learning theory". In: *Summer School on Machine Learning*. Springer. 2003, pp. 169–207.

[BBL05] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. "Theory of classification: A survey of some recent advances". In: *ESAIM: probability and statistics* 9 (2005), pp. 323–375.

[Das06] Sanjoy Dasgupta. "Coarse sample complexity bounds for active learning". In: *Advances in neural information processing systems*. 2006, pp. 235–242.

[Kää06] Matti Kääriäinen. "Active learning in the non-realizable case". In: *International Conference on Algorithmic Learning Theory*. Springer. 2006, pp. 63–77.

[BDL08] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. "Importance weighted active learning". In: *arXiv preprint arXiv:0812.4952* (2008).

[CN08] Rui M Castro and Robert D Nowak. "Minimax bounds for active learning". In: *IEEE Transactions on Information Theory* 54.5 (2008), pp. 2339–2353.

[DHM08] Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni. "A general agnostic active learning algorithm". In: *Advances in neural information processing systems*. 2008, pp. 353–360.

[BBL09] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. "Agnostic active learning". In: *Journal of Computer and System Sciences* 75.1 (2009), pp. 78–89.

[Tsy09] Alexandre Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009.

[Das11] Sanjoy Dasgupta. "The two faces of active learning." 2011.

[RR11] Maxim Raginsky and Alexander Rakhlin. "Lower bounds for passive and active learning". In: *Advances in Neural Information Processing Systems*. 2011, pp. 1026–1034.

[Duc14] John C Duchi. "Multiple Optimality Guarantees in Statistical Learning". PhD thesis. UC Berkeley, 2014.

[Han+14] Steve Hanneke et al. "Theory of disagreement-based active learning". In: *Foundations and Trends in Machine Learning* 7.2-3 (2014), pp. 131–309.

[HY15] Steve Hanneke and Liu Yang. "Minimax analysis of active learning". In: *The Journal of Machine Learning Research* 16.1 (2015), pp. 3487–3602.

[Han16] Steve Hanneke. "The optimal sample complexity of PAC learning". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1319–1333.

[WS16] Yining Wang and Aarti Singh. "Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition". In: *Thirtieth AAAI Conference on Artificial Intelligence*. 2016.

A PAC-learning

We saw in Section 2.1 that to learn a hypothesis $\widehat{h}$ that performs nearly as well as h^*, it suffices to obtain a large number of data points $(x_1, y_1), \ldots, (x_n, y_n)$ and return the empirical risk minimizer. This number n of data points required may be very large; using empirical risk minimization, we can obtain a hypothesis that has excess risk of no more than ε with probability $1 - \delta$ using a sample size n where

$$n = O\left(\frac{1}{\varepsilon^2} \log \frac{|\mathcal{H}|}{\delta}\right),$$

with respect to the underlying data distribution P_{XY}. At this point, we have placed no assumptions on our data distribution. But we might more generally assume that the data are drawn from a specific class of distributions $\mathcal{P}$. Over this class of distribution, we define *probabilistically approximately correct (PAC) learning* as being able to recover a hypothesis that is approximately correct with high probability:

Definition 33 (PAC-learning)**.** *Let $\mathcal{H}$ be a hypothesis class over $\mathcal{X}$. Let $\mathcal{P}$ be a collection of probability distributions over $\mathcal{X} \times \{0, 1\}$. We say that $\mathcal{H}$ is* PAC-learnable *over $\mathcal{P}$ if there exists an algorithm $\mathcal{A}$ such that for all $\varepsilon, \delta \in (0, 1)$, for all $\mathcal{P} \in \mathcal{P}$, by using n samples $(x_i, y_i)_{i=1}^n$, $\mathcal{A}$ returns a hypothesis*

$$\widehat{h}_n := \mathcal{A}\big((x_i, y_i)_{i=1}^n\big)$$

such that with probability $1 - \delta$,

$$R(\widehat{h}_n) \leq R(h^*) + \varepsilon.$$

We say that $\mathcal{A}$ is an algorithm that (ε, δ)-learns *$\mathcal{H}$ over $\mathcal{P}$.*

Definition 34 (Sample complexity)**.** *Let $\mathcal{X}$, $\mathcal{H}$, and $\mathbb{D}$ as before. The* sample complexity of an algorithm *$\mathcal{A}$ is the amount of data required to (ε, δ)-learn $\mathcal{H}$ over $\mathbb{D}$. We denote the sample complexity of $\mathcal{A}$ by $m(\mathcal{A}, \mathcal{H}, \mathbb{D}, \varepsilon, \delta)$. We further define $m(\mathcal{H}, \mathbb{D}, \varepsilon, \delta)$ to be the* sample complexity of (ε, δ)-learning $\mathcal{H}$ over $\mathbb{D}$ *as the minimum sample complexity of any such algorithm. When the context is clear, we denote the sample complexity by n.*

Equation 1 gives us our first sample complexity bound:

Proposition 35. *Let $\mathcal{H}$ be a hypothesis class of size N over $\mathcal{X}$. Let $\mathbb{D}$ be the collection of all probability distributions $\mathcal{P}$ over $\mathcal{X} \times \{0, 1\}$. Its sample complexity is at most:*

$$m(\mathcal{H}, \mathbb{D}, \varepsilon, \delta) = O\left(\frac{1}{\varepsilon^2} \log \frac{N}{\delta}\right). \tag{4}$$

And so, it is *information-theoretically* possible to (ε, δ)-learn $\mathcal{H}$ over any distribution by just using at most $O\left(\frac{1}{\varepsilon^2} \log \frac{N}{\delta}\right)$ i.i.d. samples. Of course, whether or not a particular algorithm can actually achieve this efficiently is a separate, though important, matter.

A.1 Infinite hypothesis classes

To extend this result to hypothesis classes of infinite size, we introduce the *Vapnik-Chervonenkis (VC) dimension* $V(\mathcal{H})$ of a hypothesis class.

Definition 36 (VC dimension)**.** *We say that* $\mathcal{H}$ shatters *the subset* $\{x_1, \dots, x_n\} \subset \mathcal{X}$ *if, for any realization* $y_1, \dots, y_n \in \{0, 1\}$*, there is some* $h \in \mathcal{H}$ *such that:*

$$h(x_i) = y_i.$$

The VC dimension*,* V*, is the largest size of a set that* $\mathcal{H}$ *shatters.*

For example, the VC dimension of a finite hypothesis class with N elements must surely be bounded above by:

$$V \leq \log_2 N.$$

Otherwise, notice that there are 2^n possible labeling for any collection of n points $x_1, \dots, x_n \in \mathcal{X}$. If $n > \log_2 N$, the number of labeling required to shatter would be more than the number of hypothesis to begin with.

In the finite case, we may in fact replace the $\log N$ term in Equation 4 with V. Additionally, it turns out that using a covering number argument coupled with a technique called chaining, this results remains true for infinite hypothesis classes too:

Theorem 37 (Passive sample complexity upper bound, agnostic)**.** *Let* $\mathcal{H}$ *be a hypothesis class with finite VC dimension* $V < \infty$*. Let* $\mathbb{D}$ *be the collection of all probability distributions* $\mathcal{P}$ *over* $\mathcal{X} \times \{0, 1\}$*. Its sample complexity is at most:*[5]

$$m(\mathcal{H}, \mathbb{D}, \varepsilon, \delta) = \widetilde{O}\left(\frac{V}{\varepsilon^2}\right). \tag{5}$$

This result holds without any assumptions made on the class of distributions $\mathbb{D}$. However, in the realizable setting, we assume that the the labels come to us deterministically—when we decompose $\mathcal{P}$ into $P_X \otimes P_{Y|X}$, the conditional probability satisfies:

$$P_{Y|X}(y|x) = \begin{cases} 1 & y = h^*(x) \\ 0 & \text{o.w.} \end{cases}$$

In this setting, we can appeal to a tighter concentration bound:

Theorem 38 (Bernstein's inequality)**.** *Let* $Z_1, \dots, Z_n$ *be* n *i.i.d. Bernoulli variables. Then:*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z]\right| > t\right] \leq 2\exp\left(-\frac{nt^2}{2\mathrm{Var}[Z] + 2t/3}\right).$$

As before, let $Z_i = \mathbf{1}[h(X_i) \neq Y_i]$ be the Bernoulli random variable corresponding to the empirical error on a single sample. But suppose that the error $\mathrm{err}(h) \ll 1$ is very small. Then, it follows that

$$\mathrm{Var}[Z] = \mathrm{err}(h)(1 - \mathrm{err}(h)) \leq \mathrm{err}(h) \ll 1$$

is also very small. In this case, Bernstein's will give an upper bound on the order of $\exp(-2nt)$; the number of samples to estimate $\mathrm{err}(h)$ up to within ε will then be $\widetilde{O}(1/\varepsilon)$. A more careful analysis implies [Han16]:

[5] The notation $\widetilde{O}(\,\cdot\,)$ is used to hide any polylogarithmic dependencies on $\frac{1}{\delta}, \log\frac{1}{\varepsilon}$, and V (i.e. terms of the form $p\left(\log\frac{1}{\delta}, \log\log\frac{1}{\varepsilon}, \log V\right)$ where p is a polynomial); they are not of great interest to us.

Theorem 39 (Passive sample complexity upper bound, realizable)**.** *Let $\mathcal{H}$ be a hypothesis class with finite VC dimension $V < \infty$. Let $\mathbb{D}$ be the collection of probability distributions for which $\mathcal{H}$ is realizable. Its sample complexity is at most:*

$$m(\mathcal{H}, \mathbb{D}, \varepsilon, \delta) = \widetilde{O}\left(\frac{V}{\varepsilon}\right). \tag{6}$$

This shows that in the realizable setting, an improvement in sample complexity by a factor of $\frac{1}{\varepsilon}$ becomes possible. Still, both sample complexities in Theorem 37 and Theorem 39 can become very large if we want a good classifier; it can be very costly to obtain so much labeled data.

In the active learning model, we'll let the learner draw unlabeled instances x and give them the option to reveal its label y. In that case, we care also about how many labels are required.

Definition 40 (Label complexity)**.** *The* label complexity of an algorithm $\mathcal{A}$ *is the amount of labeled data required to (ε, δ)-learn $\mathcal{H}$ over $\mathbb{D}$. We denote the label complexity by $m_{\text{LABEL}}(\mathcal{A}, \mathcal{H}, \mathbb{D}, \varepsilon, \delta)$. The* label complexity *of (ε, δ)-learning $\mathcal{H}$ over $\mathbb{D}$ is the minimum label complexity of any such algorithm, $m_{\text{LABEL}} := m_{\text{LABEL}}(\mathcal{H}, \mathbb{D}, \varepsilon, \delta)$.*

A.2 Covering and packing number

We recall the usual definition of an ε-cover and ε-packing of a metric space. A cover will be useful because it often allows us to closely approximate an infinite metric space with a finite subset—it will be sufficient to upper bound the difficulty of learning with respect to the cover. Likewise, a packing will help provide lower bounds on how difficult a learning problem.

Definition 41 (ε-cover, packing, and net)**.** *Let (M, d) be a (pseudo)-metric space. A subset $S \subset M$ is an* ε-cover *if for all points $x \in M$, there exists some point $s \in S$ that is ε-close to x,*

$$d(x, s) \leq \varepsilon.$$

S is a ε-packing *if all points in S are ε-separated from each other. That is, for all $s, s' \in S$,*

$$d(s, s') > \varepsilon.$$

If S is both an ε-cover and ε-packing, we say that it is an ε-net*.*

Definition 42 (Covering and packing number)**.** *Let (M, d) be a (pseudo)-metric space. The* covering number *$\mathcal{N}(\varepsilon, M, d)$ is the size of a minimal ε-cover of M. Similarly, the* packing number *$\mathcal{M}(\varepsilon, M, d)$ is the size of a maximal ε-packing of M.*

These two quantities are closely related by the following inequality:

Proposition 43. *Suppose that $\mathcal{N}(\varepsilon, M, d) < \infty$ for all $\varepsilon > 0$. Then:*

$$\mathcal{M}(2\varepsilon, M, d) \leq \mathcal{N}(\varepsilon, M, d) \leq \mathcal{M}(\varepsilon, M, d).$$

The following theorem tells us that it is possible to approximate a hypothesis class $\mathcal{H}$ with VC dimension V via a finite subset $\mathcal{H}_0$ where $\log |\mathcal{H}_0| = O\left(V \log \frac{1}{\varepsilon}\right)$.

Theorem 44 (Theorem 6, [Hau92])**.** *Let $\mathcal{H}$ have VC dimension V and let d be the induced pseudometric on $\mathcal{H}$ induced by the measure P_X. Then, for all $0 < \varepsilon \leq 1$,*

$$\mathcal{M}(\varepsilon, \mathcal{H}, d) < 2\left(\frac{2e}{\varepsilon} \ln \frac{2e}{\varepsilon}\right)^V.$$

B Technical proofs

Lemma 3 (Search implies learn)**.** *Let $\mathcal{H}$ and P_{XY} be a hypothesis class and a joint probability distribution over $\mathcal{X} \times \mathcal{Y}$, as above. Let d be the pseudometric induced on $\mathcal{H}$ by the marginal distribution P_X. If $h \in \mathcal{H}$ satisfies $d(h, h^*) < \varepsilon$, its excess risk is at most ε.*

Proof of Lemma 3. Let $g(x) = \mathbb{E}[Y|X = x]$. Then, $R(h) = d(h, g)$ and $R(h^*) = d(h^*, g)$. By triangle inequality, $d(h, g) \leq d(h, h^*) + d(h^*, g)$. Combining this with the assumption, $d(h, h^*) < \varepsilon$, we obtain:

$$d(h, g) - d(h^*, g) < \varepsilon.$$

Thus, the excess risk, $R(h) - R(h^*)$, is bounded above by ε. □

Lemma 4 (Learn implies search)**.** *Let $\mathcal{H}$, P_{XY}, and d as before. Let ν be the optimal risk of any $h \in \mathcal{H}$; that is, ν is the true risk $R(h^*)$ of a best-in-class classifier $h^* \in \mathcal{H}$:*

$$\nu := \min_{h \in \mathcal{H}} R(h).$$

If $h \in \mathcal{H}$ has excess risk at most ε, then $d(h, h^*) < 2\nu + \varepsilon$.

Proof. As before, let $g(x) = \mathbb{E}[Y|X = x]$. If h has excess risk at most ε, then $d(h, g) < d(h^*, g) + \varepsilon$. By the triangle inequality, this implies that:

$$d(h, h^*) < d(h, g) + d(g, h^*) < 2\nu + \varepsilon.$$

Thus, h is contained in the $(2\nu + \varepsilon)$-ball centered at h^*. □

Theorem 7 (A^2 label complexity)**.** *Let $\varepsilon, \delta > 0$ be fixed. Let $\theta := \theta_{\nu+\varepsilon}(\mathcal{H}, P_X, h^*)$. There exists settings for k_t and ε_t such that the A^2 algorithm (ε, δ)-learns $\mathcal{H}$ with label complexity $n = \sum_{i=1}^{T} k_t$ at most:*

$$n = O\left(\theta^2 \left(1 + \frac{\nu^2}{\varepsilon^2}\right) \log \frac{1}{\varepsilon} \log \frac{|\mathcal{H}| \log \frac{1}{\varepsilon}}{\delta}\right).$$

Proof. Set $\varepsilon_t = 2^{-t}$ and k_t as before in Equation 3. Then, after $T = \lg \frac{2}{\varepsilon}$ rounds in A^2, we've obtained a version space that is contained within $B(h^*, 2\nu + \varepsilon)$. Furthermore, in the last round, not only are the k_T labels enough to reduce the version space to fit in this ball, via Equation 2, they are enough to estimate the excess risk up to within $\varepsilon/2$. Therefore, the empirical risk minimizer at this point is ε-close to h^*. □

Lemma 13. *Let P_X be a fixed marginal distribution over $\mathcal{X}$. Let $\mathcal{H}$ be a hypothesis class that is (ρ, r, τ)-splittable. Let $0 < \Delta \leq \frac{1}{2}$ and $0 \leq \nu < \left(\frac{1}{2} - \Delta\right)\tau$. Then, $\mathcal{H}$ is $(\rho, r, \Delta)_1$-testable over the following class of distributions:*

$$\mathcal{P} = \left\{P_X \otimes P_{Y|X} : \min_{h \in \mathcal{H}} R(P_{XY}, h) \leq \nu\right\}.$$

Proof of Lemma 13. Suppose that $\mathcal{H}$ is a hypothesis class that is $(\rho, \varepsilon, \tau)$-splittable. Fix some distribution $P_{XY} \in \mathcal{P}$. By the assumption on $\mathcal{P}$, there exists some hypothesis h^* such that that $R(P_{XY}, h^*) \leq \nu$. For every collection of edge sets Q, let $S \subset \mathcal{X}$ be the collection of all ρ-splitters of Q_ε, where $(\rho, \varepsilon, \tau)$-splittability implies that the mass of S is at least τ. By the assumption on ν, because $R(h^*) \leq \left(\frac{1}{2} - \Delta\right)\tau$, the risk $R\big|_S(h^*)$ when restricted to drawing from the truncated distribution on S is at most $\left(\frac{1}{2} - \Delta\right)$, since:

$$\begin{aligned}\nu \geq R(h^*) &= P_X[S] \cdot R\big|_S(h^*) + P_X[S^c] \cdot R\big|_{S^c}(h^*) \\ &\geq \tau \cdot R\big|_S(h^*)\end{aligned}$$

Define the test T by drawing a random $x \sim P_{X|X \in S}$ from S and querying the label oracle $\textsc{label}(x)$ to obtain y to compute the functional $t_{(x,y)}(h) := \mathbf{1}\{h(x) \neq y\}$. This test using 1 sample has advantage Δ. □

Lemma 15 (Lower bound on ρ, [Das06])**.** *Let $0 < \alpha, \varepsilon < 1$. Let $\mathcal{V} \subset \mathcal{H}$. Then, $\mathcal{V}$ is $\big((1-\alpha)\varepsilon, \varepsilon, \alpha\varepsilon\big)$-splittable.*

Proof. Let $Q \subset \binom{\mathcal{V}}{2}$ be a finite edge set. Recall that the length of an edge is the probability that a random instance x will cut the edge. It follows that the expected number of edges in Q_ε cut by a random x is at least $\varepsilon \cdot |Q_\varepsilon|$. On the other hand, we have:

$$\begin{aligned}\varepsilon \cdot |Q_\varepsilon| &\leq \mathop{\mathbb{E}}_{x \sim P_X}\left[\#\text{ edges in } Q_\varepsilon \text{ cut by } x\right] \\ &\leq \Pr_{x \sim P_X}\left[x \text{ cuts more than } (1-\alpha)\varepsilon\text{-fraction of } Q_\varepsilon\right] \cdot |Q_\varepsilon| + (1-\alpha)\varepsilon \cdot |Q_\varepsilon|.\end{aligned}$$

This shows that at least an $\alpha\varepsilon$-fraction of instances will $(1-\alpha)\varepsilon$-split Q_ε. □

Theorem 16 (Splitting index label complexity)**.** *Let $\varepsilon, \delta > 0$ be fixed. Let $\mathcal{P}$ be a class of distributions for which the model error of $\mathcal{H}$ is at most ν:*

$$\max_{P_{XY}} \min_{h \in \mathcal{H}} R(P_{XY}, h) \leq \nu.$$

Let $\mathcal{H}$ have disagreement coefficient $\theta := \theta_{2\nu+\varepsilon}(\mathcal{H}, P_X, h^*)$ and splitting index $(\rho, 2\nu+\varepsilon, \tau)$. Then, there is an algorithm that (ε, δ)-learns $\mathcal{H}$ with label complexity at most:

$$n = O\left(\frac{1}{\rho}\frac{\tau^2}{(\tau-2\nu)^2}\log|\mathcal{H}| + \theta^2\left(1 + \frac{\nu^2}{\varepsilon^2}\right)\right) \cdot \log\frac{|\mathcal{H}|}{\delta}.$$

Proof. Corollary 14 shows that using at most n_1 samples, where:

$$n_1 = O\left(\frac{\tau^2}{\rho(\tau-2\nu)^2}\log|\mathcal{H}| \cdot \left(\log|\mathcal{H}| + \log\frac{1}{\delta_1}\right)\right),$$

we can obtain a version space $\mathcal{V} \subset B(h^*, 2\nu+\varepsilon) \subset \mathcal{H}$ that contains h^*. Let θ be the disagreement coefficient $\theta(\mathcal{H}, P_X, h^*)$. Once again, applying Equation 2 implies that to ε-learn, now over $\mathcal{V}$ requires at most n_2 samples, where:

$$n_2 = O\left(\theta^2\left(1 + \frac{\nu^2}{\varepsilon^2}\right)\log\frac{|\mathcal{H}|}{\delta_2}\right).$$

It follows that if we set $\delta_1 = \delta_2 = \delta/2$, combining these two procedures yields an active learning algorithm that (ε, δ)-learns with label complexity $n_1 + n_2$. □

Lemma 17. *Let $\mathcal{H}$ be the class of linear threshold classifiers on $\mathcal{X} = [0,1]$ with marginal distribution P_X. Then, the disagreement coefficient is constant $\theta_s(\mathcal{H}, P_X, h^*) \leq 2$. Furthermore, it is $\big(\frac{1}{2}, \varepsilon, \varepsilon\big)$-splittable.*

Proof. To show that the disagreement coefficient is equal to 2, note that the distance $d(h_\alpha, h_{\alpha'})$ is r (without loss of generality, assume $\alpha \leq \alpha'$) if and only if the mass of the set $(\alpha, \alpha'] \subset [0,1]$ is r. Thus, if $h_\alpha \in B(h_{\alpha^*}, r)$, this implies that α is contained in a subset of $[0,1]$ with probability mass at most $2r$. This shows that the disagreement coefficient is at most 2.

To show that $\mathcal{H}$ is $\big(\frac{1}{2}, \varepsilon, \varepsilon\big)$-splittable, consider any collection of edges n of the form: $\{h_{\alpha_i}, h_{\alpha_i'}\}$. Without loss of generality, assume that $\alpha_i + \varepsilon < \alpha_i'$ and that the α_i's are monotonically increasing:

$$\alpha_1 \leq \cdots \leq \alpha_{\lceil n/2 \rceil} \leq \cdots \leq \alpha_n.$$

Querying the label of any $x \in \big(\alpha_{\lceil n/2 \rceil}, \alpha_{\lceil n/2 \rceil}\big]$ will deterministically remove at least $n/2$ of the edges. In particular, if the label y is 0, this removes $\alpha_1, \ldots, \alpha_{\lceil n/2 \rceil}$ from the version space, reducing the number of edges by half. On the other hand, if the label is 1, this removes $\alpha'_{\lceil n/2 \rceil}, \ldots, \alpha'_n$ from the version space, since the edges have length at least ε. This again reduces the number of edges by half. □

Lemma 18. *Let $\mathcal{H}$ be the class of interval functions on $\mathcal{X} = [0,1]$, with marginal distribution P_X. Let $h^* = h_{\alpha,\beta}$, where $\gamma = \beta - \alpha$. The disagreement coefficient is:*

$$\theta_s(\mathcal{H}, P_X, h_{\alpha,\beta}) = \begin{cases} \frac{1}{s} & s > \gamma \\ 4 & s \leq \gamma. \end{cases}$$

For any $0 < s < 1$, $\mathcal{H}$ is not $(\rho, \varepsilon, \tau)$-splittable for any $\rho > 2\varepsilon$.

Proof. Suppose that $s > \gamma$. Then, any interval function $h_{\alpha',\beta'}$ with $\beta' - \alpha' < s - \gamma$ is within the $B(h^*, s)$ ball. The region of disagreement for all such interval functions is all of $\mathcal{X}$. Thus, the disagreement coefficient when $s > \gamma$ is equal to $\frac{1}{s}$.

When $s \leq \gamma$, we note that the interval (α, β) must have nonempty intersection with the interval (α', β'). For simplicity, assume P_X is the uniform distribution (otherwise, apply an order- and measure-preserving map like the CDF from $\mathcal{X}$ to $[0,1]$ with the uniform distribution). Working through the possible cases of α and β shows that the only regions of disagreement are the intervals $[\alpha - s, \alpha + s]$ and $[\beta - s, \beta + s]$. Thus, the disagreement coefficient is at most 4.

To show that $\mathcal{H}$ is not $(\rho, \varepsilon, \tau)$-splittable for any $\rho > 2\varepsilon$, we consider again the subset of hypotheses:

$$\mathcal{H}' = \left\{ h_{k\varepsilon,(k+1)\varepsilon} : 0 \leq k < \frac{1}{\varepsilon} \right\}.$$

The complete graph on $\mathcal{H}'$ has at least $\left\lfloor \frac{1}{\varepsilon} \right\rfloor^2 / 2$ edges of length greater than ε. Yet, any label will potentially remove only a single hypothesis, and with it, at most $\frac{1}{\varepsilon}$ edges. Thus, $\mathcal{H}$ is not $(\rho, \varepsilon, \tau)$-splittable for $\rho > 2\varepsilon$. □

Proposition 26 (Theorem 12, [BDL08])**.** *Fix $0 \leq \nu \leq \frac{1}{2}$. Let $\mathcal{H}$ be a hypothesis class with VC dimension V. There exists a class of distributions $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ such that $R(h^*) = \nu$ and any algorithm that (ε, δ)-learns $\mathcal{H}$ over $\mathcal{P}$ must have label complexity at least:*

$$n = \Omega\left(V \cdot \frac{\nu^2}{\varepsilon^2} \right).$$

Proof. Let $x_1, \ldots, x_V \in \mathcal{X}$ be shattered by $\mathcal{H}$. We'll construct a class of distributions such that the first $d-1$ points are noisy and have an α-fraction of the probability mass while the remaining point has a $(1-\alpha)$-fraction of the mass and is noiseless. Let the noise rate on the noisy points be $\frac{1}{2} - \beta$. It follows if we set $\alpha\beta = \varepsilon$, then any ε-good hypothesis must be correct on at least half of the noisy points.

On the other hand, we should set the average noise rate to be $\nu = \alpha\left(\frac{1}{2} - \beta\right)$. It follows that we can set $\alpha = 2(\nu + \varepsilon)$, where the noise is split evenly across these points; the margin β for each point is:

$$\beta = \frac{\varepsilon}{2(\nu + \varepsilon)}.$$

Let's set β to a value so that any algorithm that ε-learns $\mathcal{H}$ must successfully perform binary hypothesis tests on a constant fraction of the noisy points. In particular, let's set $\alpha\beta = \varepsilon$, that way the learner must correctly label half of the noisy points. From the binary hypothesis testing lower bound, to correctly label a single noisy point requires $\Omega\left(\frac{1}{\beta^2}\right)$ samples. It follows that to correctly label $\left\lfloor \frac{V}{2} \right\rfloor$ of the points, an ε-learner must have label complexity:

$$n = \Omega\left(V \cdot \frac{\nu^2}{\varepsilon^2} \right).$$

□

Lemma 28 (KL-divergence of sequentially sampled data, [RR11])**.** *Let S_m be a m-step learning strategy and let $P_{Y|X,\theta'}$ and $P_{Y|X,\theta}$ be two conditional distributions. Suppose that they induce the sample distributions $\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}$ and $\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta}}$. Then:*

$$\mathrm{KL}\left(\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}\middle\|\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta}}\right)=\sum_{x\in\mathcal{X}}N_{\theta'}(x)\cdot\mathrm{KL}\left(P_{Y|X=x,\theta'}\middle\|P_{Y|X=x,\theta}\right),$$

where $N_{\theta'}(x)=\mathbb{E}_{\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}}\left[\#\{t:x_t=x\}\right]$ is the expected number of times x will be queried when given responses from $P_{Y|X,\theta'}$.

Proof.

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}\middle\|\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta}}\right)&=\sum_{x^m,y^m}\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}(x^m,y^m)\log\frac{\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}}{\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta}}}\\
&=\sum_{x^m,y^m}\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}(x^m,y^m)\sum_{i=1}^m\log\frac{P_{Y|X,\theta'}(y_i|x_i)}{P_{Y|X,\theta}(y_i|x_i)}\\
&=\sum_{i=1}^m\sum_{x,y}\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}(X_i=x,Y_i=y)\log\frac{P_{Y|X,\theta'}(y_i|x_i)}{P_{Y|X,\theta}(y_i|x_i)}\\
&=\sum_{i=1}^m\sum_{x}\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}(X_i=x)\mathrm{KL}\left(P_{Y|X=x,\theta'}\middle\|P_{Y|X=x,\theta}\right)\\
&=\sum_{x\in\mathcal{X}}\sum_{i=1}^m\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}(X_i=x)\mathrm{KL}\left(P_{Y|X=x,\theta'}\middle\|P_{Y|X=x,\theta}\right)\\
&=\sum_{x\in\mathcal{X}}N_{\theta'}(x)\mathrm{KL}\left(P_{Y|X=x,\theta'}\middle\|P_{Y|X=x,\theta}\right)
\end{aligned}$$

Note that the first step is taken by expanding the second term (blue):

$$\begin{aligned}
\log\frac{\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}}{\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta}}}&=\log\frac{\prod_{i=1}^m\Pi^{(i)}_{X_i|X^{i-1},Y^{i-1}}(x_i|x^{i-1},y^{i-1})P_{Y|X,\theta'}(y_i|x_i)}{\prod_{i=1}^m\Pi^{(i)}_{X_i|X^{i-1},Y^{i-1}}(x_i|x^{i-1},y^{i-1})P_{Y|X,\theta}(y_i|x_i)}\\
&=\sum_{i=1}^m\log\frac{P_{Y|X,\theta'}(y_i|x_i)}{P_{Y|X,\theta}(y_i|x_i)}.
\end{aligned}$$

Also, in the third step, we rely on the property that $\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}(X_i,Y_i)$ is a product distribution:

$$\sum_{x,y}\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}(X_i=x,Y_i=y)=\sum_x\mathbb{P}^{\mathsf{S}_m,P_{Y|X,\theta'}}(X_i=x)\sum_y P_{Y|X,\theta'}(Y_i=y)$$

□