Learning from Noise:

Applying Sample Complexity for Political Science Research

Perry Carter[*] Dahyun Choi[†]

August 6, 2024

While statistical learning bridges the gap between theoretical concepts and complex empirical realities, the question of what constitutes “good enough" data for social scientists remains understudied. In this article, we introduce the *Probably Approximately Correct* model and present *sample complexity* bounds, which take advantage of researchers-specified estimates of labeling error to guarantee the sample size required for a minimum level of accuracy. We develop a simulation-based approach to demonstrate its feasibility and provide the `scR` R package, offering a computationally efficient way to implement the proposed methods. We aim to improve standard practice by providing a general-purpose tool to validate the quality of measures when fuzzy measurement boundaries make generating data with ground-truth labels infeasible.

Keywords: Probably Approximately Correct Model, Sample Complexity Bounds, Vapnik-Chervonenkis Dimension, Measurement, Sample Size
Word Count: 7615

[*] Ph.D. Candidate, Department of Politics, Princeton University.Email: pjcarter@princeton.edu

[†] Ph.D. Candidate, Department of Politics, Princeton University. Email: dahyunc@princeton.edu

1. Sample Size Calculations in Applied Research

What constitutes “good enough" data? The majority of concepts of interest in political science are not directly observable; consequently, many scholars design and construct their own measurements from messy or unstructured sources of data. There has been extensive work on the use of hand-coding or crowdsourced annotations (e.g., Tian and Zhu 2015; Benoit et al. 2016; Carlson and Montgomery 2017; Ying, Montgomery, and Stewart 2022; Miller, Linder, and Mebane 2018), and researchers have attempted to improve the efficiency of measurement by using machine learning to extrapolate from a relatively small training set to a larger set of unlabeled or unseen data (e.g., Grimmer and Stewart 2013; Barberá et al. 2021; Jerzak, King, and Strezhnev 2023).

However, despite the explosion in the use of machine learning, the issue of how data affect measurement quality and inference has received relatively little attention in the political science literature. While measurement models are typically optimized for accuracy on observed labeled instances, the ability to make *generalizable* claims beyond the training data is the ultimate goal of statistical learning. Yet, if the training data is not representative of the population of interest, the model may not generalize well to other samples or populations, often due to biased sampling methods. Similarly, high variability within the data can make it challenging for researchers to quantify the resources needed to “learn” from given samples or to determine whether learning from such noisy data is always possible.

Relatedly, one typical implicit assumption in most applications is that learning algorithms have access to noise-free ground-truth labels[1] for training examples of the target concept[2]. However, this assumption can further expose empirical analysis in

[1] That is, a mechanism generating accurately labeled examples from a known probability distribution.

[2] We use the term *concept* in this paper to refer to a latent binary class. We prefer this term over the equivalent “class” or “latent variable” in order to highlight the close relationship to concept formation problems in social science. A technically precise definition is provided in Section 3.

political science to the fragility of statistical inference, given that many concepts in the discipline such as transparency (e.g., Hollyer, Rosendorff, and Vreeland 2014), electoral competitiveness (e.g., Kayser and Lindstädt 2015; Cox, Fiva, and Smith 2020), or conflict initiation (e.g., Esarey and Pierce 2012) are subject to conceptual ambiguity or "stretching" (Collier and Mahon 1993). Such latent concepts are typically both high-dimensional and have ambiguous boundaries, making it difficult to specify explicit conditions for inclusion *a priori*. However, despite the inherent complexity of measurement tasks in social science, "learning" from the noisiness inherent in quantitative political science research has rarely been addressed.

To address this lacuna, we propose the application of an approach based on the *Probably Approximately Correct* (PAC) model, which uses the training set to learn a concept that has a small true error on the test distribution.[3] The premise of the PAC model is that examples are drawn from a fixed, but unknown, distribution over the instance space. This assumption provides researchers with hope that what they learn from the training data will generalize to new and unseen test data. Based on this setup, researchers can then determine the sample size necessary to guarantee a probably approximately correct solution.

The smallest sample size necessary to learn a concept is referred to as *sample complexity* (Blumer et al. 1989). Simply put, the sample complexity bound guarantees the sample size required to achieve a minimum level of accuracy with a precise level of confidence, for all distributions and all target concepts. Therefore, it allows researchers to infer the generalizability of the concepts of interest beyond the training set they use. To determine the sample complexity bound, we incorporate the estimation of the Vapnik-Chervonenkis (VC) dimension, or "capacity," of the target concept. The method then allows researchers to theoretically evaluate how much data is needed for their

[3] PAC learning assumes that labeled instances are coming from a fixed but unknown distribution and that there is an unknown concept belonging to a known class that truly labels the instances. We will provide further details in the section on PAC learning.

design before making costly investments in data collection. In addition, we provide an illustrative example of how the method can be extended to hypothesis classes with infinite VC dimensions by applying the union bound to countably infinite unions of concepts.

After introducing the notion of the PAC model and sample complexity, we provide a novel simulation-based approach to compare the empirical performance of sample complexity bounds for researcher-specified confidence, accuracy, and misclassification parameters. We further demonstrate the feasibility of this approach, implemented in a companion *R* package. A simulation exercise using a foundational concept in political science, *polyarchy*, verifies that the sample complexity bound we produce is not overly conservative. To further validate the usefulness of the method, we provide a replication analysis of Dressel and Farid (2018) and show how the proposed method can be used for *a priori* analysis before researchers engage in data collection. Given that the collection of human-labeled data tends to be expensive and time-consuming, we anticipate that the application of sample complexity will offer significant time and cost-saving advantages to researchers.

This paper aims to provide a general-purpose tool for assessing the sample size needed to achieve adequate performance from statistical learning algorithms in applied social science. This provides a foundation for the supervised learning tasks prevalent in political science using a principled approach analogous to the hypothesis testing framework for statistical inference. Moreover, when researchers intend to estimate causal effects of or on latent concepts as a downstream application, as in the persistent debate over the causal effects of democracy (Acemoglu et al. 2019), measurement error has direct implications for the power of the corresponding hypothesis test (Knox, Lucas, and Cho 2022). By providing a straightforward way to assess the consequences of sample size for measurement, we therefore offer a means to improve the accuracy of power analysis for causal inference.

2. Principled and Unprincipled Measurement-Inference in Social Science

Significant progress has been made in the application of machine learning within political science. According to Arnold et al. (2023), 64 manuscripts relevant to machine learning were published in three leading political science journals, including *American Political Science Review* , *Political Analysis,* and *Political Science Research and Methods,* between 2016 and 2021. Scholars have improved machine learning models through regularization (e.g., Hainmueller and Hazlett 2014; Fariss and Jones 2018) or parameter tuning (e.g., Neunhoeffer and Sternberg 2019) to produce a good indicator of prediction in new data. These advances in machine learning techniques have naturally been accompanied by an increase in inferential tasks (e.g., Grimmer, Westwood, and Messing 2014).

However, despite the abundant interest given to improved performance in machine learning models, there has been little attention to downstream questions concerning “measurement-inference" (Grimmer 2015). The question of what constitutes sufficient data for feasible statistical inference is a principled one that has been recognized by most scholars using machine learning techniques (e.g., Hopkins and King 2010). Yet the answer to this question has been unprincipled, as researchers have designed their studies within the arbitrary scopes allowed by their resources and efficiency.[4] For decades, there has been a lack of formalized justification for determining the sufficient sample size needed for generalizable measurement, taking into account the degree of accuracy a researcher can expect from such samples. Our review of publications in *Political Analysis* for 2023 that employ applied machine learning methods reveals that

[4] In fact, Hopkins and King (2010) presents the average root mean square error by varying sample sizes, concluding that “more than about 500 documents to estimate a specific quantity of interest is probably unnecessary," due to inefficiency.

none of them have attempted to justify whether “learning" has been successful given the sample sizes chosen in their quantitative empirical studies. While some have discussed train-test splits within the scope of accessible data, the consideration of measuring complex concepts in political science and the feasibility of learning amidst noisy data has not been addressed.

Principled accounts of what constitutes “good enough" data are crucial not only for helping researchers identify the conditions under which meaningful “learning" is possible but also for better testing of causal theories involving unobservable variables. Researchers often map messy and high-dimensional data to low-dimensional measures to capture latent concepts in political science and to empirically test causal theories. However, these imperfect and noisy measures can produce biased point estimates and standard errors (Knox, Lucas, and Cho 2022), as standard practices implicitly ignore that what is "learned" from the data is also estimated from the subset of data available to researchers, instead simply assuming perfectly observed latent constructs. An active body of work aimed at addressing these issues, arising from measurement uncertainties, is through a design-based approach (e.g., Wang, McCormick, and Leek 2020; Fong and Tyler 2021; Egami et al. 2024). This approach seeks to recover valid standard errors and confidence intervals in consideration of unknown non-random prediction errors.

While design-based approaches provide powerful tools for the researcher to mitigate the bias induced by noise in measurements, the model still relies on the accuracy of predictions on sampled data, which may not provide an accurate picture of the models' predictive generalizability. We directly address this issue and consider the issue of "generalizable" inference from downstream data when there is a probability that the training examples are non-representative or incorrectly labeled. Our proposed approach of *sample complexity* bounds offers performance guarantees for a model based on a given number of samples, providing more rigorous insights concerning the measurement challenges in social science than empirical prediction error estimates. In the following section, we present a formal framework for statistical learning before discussing the proposed approach.

An important point of clarification regards our framing of the problem as one of identifying what constitutes *good enough* data. In political methodology – particularly from the Bayesian perspective – it is standard to take the available data as given and to focus instead on designing an estimator that has desirable properties for a given research context. In this paper, we take the opposite perspective, assuming that a valid estimation strategy corresponding to the concept being measured has been selected but that the researcher is at the *design* stage of a research project, prior to collecting data.

Hence, the question faced by practitioners is conceptually equivalent to that of power analysis at the pre-analysis stage of experimental research: given access to *infinite* data with ground-truth labels, the algorithm is known (or at least assumed) to be able to perform arbitrarily well. Due to real-world resource constraints, however, the researcher seeks to identify how much data and what degree of labeling accuracy is *sufficient* to guarantee a desired level of performance before investing in data collection. As in power analysis, where the target power level is determined by the researcher's individual risk tolerance for Type II error, the accuracy and confidence parameters in our framework our chosen based on the ultimate goal, whether downstream inference or pure description.

3. A Framework for Statistical Learning

Before discussing the PAC model and sample complexity bounds, we describe a formal framework designed to capture statistical learning tasks. To link the model more closely to its practical applications, we consider a stylized version of the well-known model of "polyarchy" proposed in Dahl (2008). Here, the goal is to learn where the rectangular cutoff between polyarchies and non-polyarchies lies in a region defined by two latent dimensions – "participation" and "contestation". While we thus abstract away from much of the discussion of measurement that has dominated the recent literature on this topic (e.g., Little and Meng 2023), the basic issue of determining the cut-off point between democracies and nondemocracies remains highly salient both theoretically and empirically (Baltz, Vasselai, and Hicken 2022) and is complicated by small sample size. We therefore employ this as a running example to illustrate the approach. In the basic statistical learning setting, researchers have access to the following (Laird 2012):

Domain set: This refers to an arbitrary set, $\mathcal{X}$. This is the set of all observations that researchers might hope to label. For example, regarding the polyarchy problem just outlined, the domain set would be the set of all political regimes. These domain points are represented by a vector of features. In our case, this is a two-dimensional vector of *contestation* and *participation*. In the following discussion, we interchangeably refer to domain points as *instances* and χ as instance space.

Label set: For our discussion, we restrict the label set to be a binary classification task, with $\mathcal{Y}$ denoting the set of possible labels. For instance, let $\mathcal{Y} = \{0, 1\}$, where 1 represents polyarchies and 0 stands for non-polyarchies.

Training data: $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ indicates a finite sequence of pairs in $\mathcal{X} \times \mathcal{Y}$: that is, a sequence of labeled domain points. This is the input that researchers have access to, such as a set of countries and their observable attributes. We interchangeably describe such labeled examples S as training examples or training sets.

The learner's rule: The learner is expected to produce a prediction *rule,* which is the concept researchers would like to learn. It is a function h: $\mathcal{X} \to \mathcal{Y}$. In computational learning theory, this function is also described as a *predictor,* a *hypothesis,* or a *classifier*. The classifier can be used to predict the label of new domain points. In our polyarchy example, it is a rule that our learner will use to predict whether future countries it examines are polyarchies or not. We use the notation h to denote the hypothesis that a learning algorithm produces when given the training set.

Data-generating Process: We now discuss how the training set is generated. It is worthwhile to note that the instances – the countries that researchers could access – are generated by some arbitrary fixed probability distribution. We denote the probability distribution over $\mathcal{X}$ by D. To clarify, we do not assume that the learner has any priors about this distribution, or even that is restricted to a particular class. Each pair in the training sets S is generated by first sampling a point x_i according to D and then labeling it by $f\colon \mathcal{X} \to \mathcal{Y}$.

Measures of errors: The error of a classifier represents the probability that it does not predict the correct label on a random data point generated by the *same* underlying distribution as the training data. In other words, the error of h is the probability of drawing a random instance x according to the distribution D such that $h(x) \neq f(x)$. Given that there is an unknown concept c which determines the true label of instances, the set of labeled instances $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ is generated by taking $x_i \sim D$ i.i.d and observing the corresponding $y_i = f(x_i)$. Then the true error and empirical error can be defined as follows.

DEFINITION 1. ***True Error:*** *Consider a data-generating distribution D and the true labeling concept c. The true error of a concept h with respect to D is the probability that h makes a mistake.*

$$R(h) = Pr_{x\sim D}[h(x) \neq y] \tag{1}$$

DEFINITION 2. ***Empirical Error:*** *Given a sample set S, the empirical error of a concept h with respect to S is the fraction of instances in S that are incorrectly labeled by h.*

$$\hat{R}_m(h) = \frac{1}{m}\sum_{i=1}^{m} 1(h(x_i) \neq y_i) \tag{2}$$

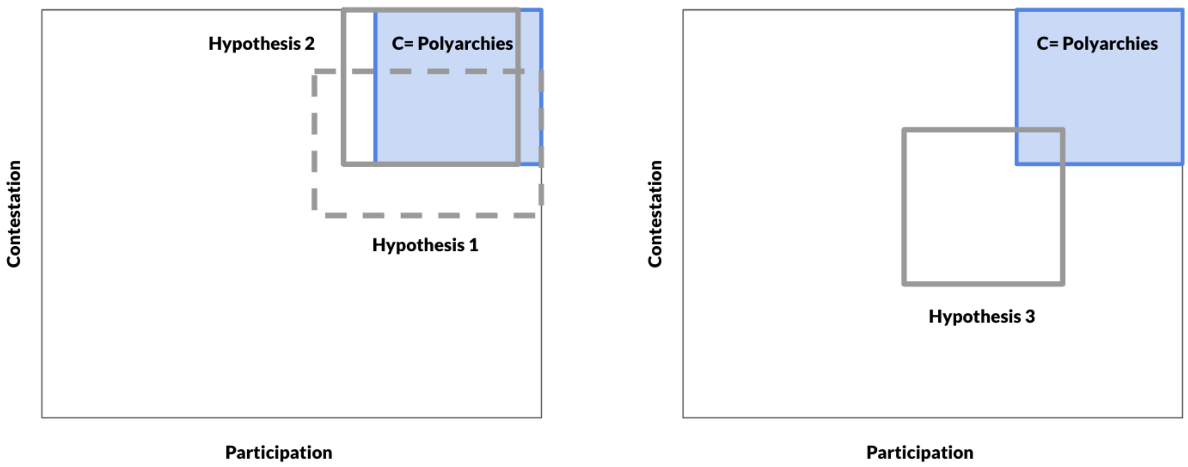
The true error of h is also referred to as generalization error or risk, which are used interchangeably throughout this article. Likewise, the terms empirical error and empirical risk are often used interchangeably. We again emphasize that the learner is blind to the underlying distribution D over the world and to the labeling function f. The learner interacts with the environment solely by observing the training set. Therefore, learning is “distribution-free" in the sense that the probability distribution generating the samples can be any distribution on the underlying measurable space (Vidyasagar 2013). As such, there is a complete absence of prior knowledge about the underlying distribution and the bounds are intended to function even in this setting. While this assumption might seem extreme, the fact that statistical learning can operate without reliance on specific distributions enables us to establish universal conditions needed for a concept to be learned.

The advantage of this general learning framework marks a significant departure from traditional methods. Typically, empirical analysis in political science relies on specific assumptions or predefined functional forms, such as linear relationships among covariates or data generated from normal distributions. In contrast, PAC learning takes a distribution-free framework whereby researchers assume that samples are created by drawing points independently at random according to an unknown fixed probability distribution and classified based on a target concept. It therefore seeks to identify models that most accurately represent the data-generating process while minimizing assumptions about the distribution.

4. Probably Approximately Correct (PAC) Learning

We begin this section with an intuitive explanation of PAC learning, followed by its formal definition.

FIGURE 1. Schematic Illustration of Learning Polyarchies from PAC Learning



In this example, the instance domain consists of a set of regimes, and the concept c we would like to learn is the distinction between polyarchies and non-polyarchies. The sample distribution is the known distribution of countries. In general, each researcher may have a different training set, and we do not know which one learns the concept. In the plot on the left-hand side of Figure 1, both hypotheses 1 and 2 are consistent; however, hypothesis 2 learns the concept better, implying that it has a smaller true error. Therefore, we hope that our error based on the training samples is smaller than some number ϵ, which we choose to define our desired level of accuracy. At this stage, our approach is considered to be “***approximately correct***".

However, since the training set researchers use is randomly drawn, there is always a nonnull probability that the drawn sample contains misleading instances, as shown in the right-hand plot of Figure 1. Therefore, we can never be 100% certain that the true error of our classifier will not exceed ϵ. Instead, we bound the probability of this occurring with δ, our confidence parameter. Thus, our model is now “***probably approximately correct***".

In sum, the accuracy parameter ϵ determines how close the output can be to the optimum and the confidence parameter δ indicates the likelihood that the classifier will meet the accuracy requirement. Under empirical settings, these approximations are inevitable. There is always a small chance that the examples that are drawn will happen to be noninformative. For example, there is always some chance that the training set will contain only one domain point, sampled over and over again as the training set is randomly generated. Furthermore, even if we are fortunate enough to obtain a training sample that closely approximates the true underlying distribution, it may still miss some fine details of the target class. In this regard, the confidence parameter addresses the error-prone nature of classification tasks.

The formal intuition behind “learning" is quite straightforward. Given a hypothesis class, H, the learner evaluates the risk, $|R(h) - \hat{R}_S(h)|$, of each h in H on the given sample and outputs a member of H that minimizes the empirical risk. The goal is that an h that minimizes the empirical risk on the sample S also minimizes the risk or has a risk close to the minimum, for the true data probability distribution. To achieve this, it’s enough to ensure that the empirical risks of all members of H closely approximate their true risks. In other words, we need the empirical risk to be uniformly close to the true risk across all hypotheses in the hypothesis class.

DEFINITION 3. ***ϵ-representative:*** *A training set S is ϵ-representative for domain $\mathcal{X}$, hypothesis class $\mathcal{H}$, and distribution $\mathcal{D}$ if*

$$\forall h \in H, |R(h) - \hat{R}_S(h)| < \epsilon \tag{3}$$

Lemma C1 further states that whenever the sample is $\frac{\epsilon}{2}$-representative, the learning rule is guaranteed to return a good hypothesis. The details and proofs can be found in the Appendix. This lemma indicates that, to ensure that the minimizing empirical risks rule functions as an agnostic PAC learner, it is enough to prove that with a probability of at least $1 - \delta$, the randomly chosen training set will be an ϵ-representative training set. In the study of machine learning theory, such property has been described as a uniform convergence property. The term “uniform" here refers to having a fixed sample size that works for all members of H and over all possible probability distributions over the domain. A formal definition is provided in Section C in the Appendix.

Based on the intuition of the PAC model, the next section further offers formal explanations the sample complexity which is a function of accuracy, ϵ, and confidence, δ, parameters.

4.1. A Sample Complexity Bound For Applied Research

For a concept to qualify as PAC-learnable, it must satisfy guarantees for all possible target concepts and across all distributions. In this section, we examine the number of training samples required to ensure the feasibility of PAC learning, referred to as *sample complexity*. To reflect the reality where a set of labeled examples provided to the learning algorithm are often "noisy," we consider the scenario where each example received by the learner is mislabeled randomly and independently with a fixed probability, $\eta < \frac{1}{2}$. Sections A and B in Appendix shows the contrasts between noise-free and noisy learning in further detail.

We therefore seek to determine the minimum sample size that produces a hypothesis within a specified error tolerance of the true concept with high probability when the data used for learning is corrupted by noise. To achieve this, the learning algorithm should be provided with a number N of i.i.d. training examples along with corresponding correct classifications. To be more precise, we define the sample complexity as the minimal integer that satisfies the requirements of PAC Learning for a given ϵ and δ. This concept can be thought of as analogous to the widely used approach of power analysis in experimental research, as both provide a measure of the sample size needed to guarantee a certain level of accuracy. As we will show, the sample complexity also has direct implications for determining power when researchers aim to identify the causal effects of a latent concept. However, sample complexity differs from power analysis in the experimental or quasi-experimental setting. Power analysis assumes certain distributional features – typically encoded as an effect size – and yields an explicit likelihood that researchers will correctly reject a hypothesis with N observations. Sample complexity, on the other hand, provides a bound on the minimum number of observations needed to achieve a target accuracy with given confidence across all distributions and target concepts.

Formally, the *sample complexity* is the minimum value of N for which the equation (3) holds. Based on the discussion above, we consider a probability distribution D over $\mathcal{X}$. We assume that the instances we observe are independent and identically distributed (i.i.d) according to an unknown D. Given that there is an unknown concept c which determines the true label of instances, the set of labeled instances $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ is generated by taking $x_i \sim D$ i.i.d and observing the corresponding $y_i = c(x_i)$. Suppose we have a model that produces a hypothesis $h \in \mathcal{H}$, given a sample of N training examples. The algorithm is called *consistent* if for every ϵ and δ, there exists a positive number of training examples $N \in \mathbb{N}$ such that for any distribution p^*, it holds that $|R(h) - \hat{R}_S(h)| \leq \epsilon$.

For simple and geometrically regular concepts, it is often possible to exploit known features of the boundary's shape to recover bounds that are quite tight. In the polyarchy example, the classification boundary is known to take the form of a rectangle with one vertex located at the point $(1, 1)$. Suppose that our learning algorithm then draws the tightest possible rectangle that encloses all positive examples seen in the training data. We take the regime that displays the lowest levels of contestation and participation while still being labeled (for now, let us assume correctly) as a polyarchy and denote it as the lower-left vertex of the bounding rectangle.

By design, this algorithm will then correctly classify all non-polyarchies as such, since it will never observe a polyarchy that falls outside of the true boundary C. In other words, it produces only Type-II error. Now fix a desired error rate ϵ and consider a hypothesized boundary h with a true error rate $R(h)$ of at least ϵ. Then it must be the case that the lower-left-most observed polyarchy was sufficiently far from the boundary as to leave at least ϵ probability of observing polyarchies with less contestation *or* participation. In other words, we can draw rectangles r_1 and r_2 from the left and bottom boundaries of C such that the probability of drawing an observation $x \in r_1$ or $x \in r_2$ is at least $\epsilon/2$ and h must disagree entirely with at least one of them. Hence we can conclude that

$$\begin{aligned} P(R(h) > \epsilon) &\leq \bigcup_{i=1,2} P(x \notin r_i) \\ &\leq 2\left(1 - \frac{\epsilon}{2}\right)^m \\ &\leq 2e^{\frac{-m\epsilon}{2}} \end{aligned}$$

Now recall that we are interested in bounding the probability of this occurring – that is, the probability that we do not draw such observations – by δ. Hence we have that $\delta \geq 2e^{\frac{-m\epsilon}{2}}$ or, equivalently,

$$m \geq \frac{2}{\epsilon} \ln\left(\frac{2}{\delta}\right) \tag{4}$$

so that in order to have an error rate below 1% with 95% confidence, we would need at least 738 observations.

This bound relies on the particular rectangular shape of our toy example and cannot be applied more generally. However, the logic used to derive it highlights an important feature of the more general bound we state below: it is entirely *distribution-free*. The distribution-free property of bounds is a double-edged sword for researchers. The primary advantage is that these bounds are generally applicable since they do not assume a specific form for the data distribution, extending their use to a wide range of problems and datasets. Moreover, they guarantee validity regardless of the distribution, making them more reliable in practice. This frees researchers from concerns about misspecification, which can lead to suboptimal performance if the assumed distribution does not match reality.

However, these bounds can often be conservative, accounting for all possible distributions. For instance, if the data are known to be drawn from best-case distribution in which all but ζ weight is applied to a set of cardinality $|\mathcal{X}| = VC(\mathcal{H})$ that are arbitrarily close to the classification boundary will yield the same sample complexity as one in which all but ζ weight is assigned to observations that are infinitely far from the classification boundary. This is because of the assumption that the test data comes from the same distribution as the training data: while the former model will learn a concept that is much closer to the true one in terms of metric distance, its risk is no lower since the test examples will also be very close to the boundary. If this assumption is violated, it may be possible to do much better – or worse – than the sample complexity bounds. For instance, if the test set is a uniform distribution over the entire feature space, but researchers are able to construct a training set out of observations that are known *ex ante* to be the most informative, then high accuracy can be achieved with fewer examples than (4) would suggest.

As a result, our bounds are potentially less useful when researchers have some prior knowledge about the data distribution which could lead to tighter bounds or better performance. Nonetheless, the simulation exercise presented in Section 5 alleviates this concern by confirming that our sample complexity bound is not overly conservative for typical applications.

We now provide a general lower bound on sample complexity given a pre-specified error rate, which researchers can utilize in circumstances where obtaining noise-free labels is challenging. Again, we consider the following scenario using the same notation as used above. Researchers have a class of functions including the target h, and accuracy and confidence parameters, ϵ, δ. Then, classifiers gain information about the target function by viewing examples labeled by f, subject to i.i.d. misclassification probability $\eta < 0.5$, and attempts to learn the target concept c according to a generic algorithm A. The goal is to generate outputs with out-of-sample accuracy of at least $1 - \epsilon$ with confidence at least $1 - \delta$; that is, for an algorithm A producing hypotheses h with error rate $e = |R(h) - \hat{R}(h)|$, we seek to achieve $P_A(e > \epsilon) \leq \delta$. While Section 3 provides a comprehensive overview of the components of statistical learning, Table 1 offers a summary of notations to help readers follow the discussion on sample complexity.

TABLE 1. Notations Used for Sample Complexity Bounds

Notations	**Explanations**
ϵ	Accuracy parameter
δ	Confidence parameter
η	Misclassification rate (= noise rate)
e_N	Error rate with N examples (= $\lvert R(h) - \hat{R}(h)\rvert$)
f	Classifier used to generate labels
c	true underlying concept
Ω	Lower bound of sample complexity (= best case complexity of an algorithm)
VC(c)	Vapnik–Chervonenkis (VC) dimension (= underlying complexity of target concepts)

We would like to note that determining sample complexity is an active research area, and numerous variants of these bounds have been introduced.[5] Therefore, we use

[5] For example, there have been discussions among scholars in machine learning theory about whether

the most widely accepted and standard bound acknowledged in the field of statistical learning. As provided and Simon (1993) with Aslam and Decatur (1996), a general lower bound on sample complexity (SCB) is given by

$$\min\left(N : P(e_N > \epsilon) < \delta\right) = \Omega\left(\frac{VC(c)}{\epsilon(1-2\eta)^2} + \frac{l\,og(1/\delta)}{\epsilon(1-2\eta)^2}\right) \tag{5}$$

where $VC(F)$ indicates the Vapnik–Chervonenkis (VC) dimension, which is a measure of the *capacity* – that is, the underlying complexity – of the target concept[6]. Note that the Ω notation is used here in the sense introduced by Knuth (1976) to imply the inverse of big-O; that is, $f = \Omega(g) \Leftrightarrow g = O(f)$. This common variation on big-Oh asymptotic notation is used in much of the literature on learning theory, and represents a lower bound in the same sense that big-Oh represents an upper bound.

Returning to our running example, a researcher planning to classify regimes into polyarchies would first need to determine the misclassification rate η. In this context, this could be thought of as the residual error of existing democracy measures, which is assumed to be i.i.d. across countries. Next, she would determine a target accuracy $1-\epsilon$ with which they want to be able to classify polyarchies not in the training set (for instance, from new historical data), and a confidence level $1-\delta$ with which she wishes to achieve this accuracy. We note that the residual δ probability of misclassifying more than ϵ proportion of polyarchies need not necessarily imply accuracy *much* worse than ϵ, but simply that it exceeds it. Finally, they would need an accurate estimate of the VC dimension of the concept of polyarchy, and an estimation algorithm capable of learning

the logarithmic factor can be removed for a particular set of classifiers (e.g., Hanneke 2016). While we stay away from this discussion, the bound we present is more conservative than the bounds where logarithmic factors are removed. Therefore, the bound provided in Equation 5 is sufficient for PAC Learning in any case.

[6] Although mismatch between measurement and concept is a potentially serious concern in applied research, we will abstract away from this issue by assuming that the algorithm being used is capable of perfectly learning the target concept given infinite correctly-labeled data, so that we can substitute the unknown $VC(F)$ with the known $VC(A)$ without loss.

it.

In this instance, since we assume in our stylized example that the classification boundary for polyarchies is simply a rectangle in *contestation–participation* space, it has a known VC dimension of 4[7]. Suppose then that the researcher has determined values of $\eta = 0.05$, $\delta = 0.05$, $\epsilon = 0.01$, indicating that she wishes to achieve 99% accuracy with 95% confidence, and believes there to be 5% of regimes in the available training data that are misclassified. Then these values can be substituted into (5), leading to the conclusion that a minimum of 864 examples would be needed to learn the concept of polyarchy – significantly exceeding the currently extant number of regimes. Hence, the researcher would discover that she needed to either find an additional source of data or to revise her accuracy target.

It is worthwhile to note a few interesting properties of sample complexity bound before examining its validity using simulation-based approaches. The conventional statistical approach used by social scientists often focuses on asymptotic properties, such as the convergence of sample-based statistical estimates as the sample size increases. In contrast, the core idea of sample complexity bounds is concerned with finite-sample bounds. This means that when researchers are given a specific sample size, we aim to capture the expected degree of accuracy based on samples of fixed size. For example, if a political scientist hypothesizes that a certain type of political parties or elections might correlate with democratic backsliding, they would check a sample of countries and test the validity of their guess through hypothesis testing, invoking asymptotic convergence to generate confidence intervals. However, when measuring the *concept* of democratic backsliding, we use sample data to identify meaningful patterns that may not have been detected by traditional observers. Sample complexity provides the necessary sample

[7] In the preceding discussion, we applied the prior knowledge that the rectangle is anchored at one corner, reducing the VC dimension to 2. Here, we apply the more general bound from the perspective of a researcher who knows the class to which the true concept belongs, but has no further information about its location in feature space.

size for researchers to determine whether these patterns are significant in the sense of yielding high predictive accuracy.

The application of sample complexity requires that for a large enough sample size, we get a hypothesis with arbitrarily small error with arbitrarily high probability, no matter what concept in H we are trying to learn or what distribution D it is drawn from. Therefore, the bounds on the sample size must be independent of the underlying distribution, the fixed probability distribution P. A class of concepts with a learning function that satisfies this condition is called *uniformly learnable*. It can be formalized by requiring that the hypothesis has an error greater than ϵ with probability at most δ for small ϵ and δ, uniformly for all concepts. The smallest size that achieves this for all distributions and all target concepts is called the complexity of the learning function. In other words, for any distribution and target concepts, with probability at least $1 - \delta$, learning algorithms produce with error at most ϵ.

This general definition of uniform learnability implies that sample sizes are uniform with respect to labeling rules and the underlying distribution. Moreover, it is also worth highlighting that the bounds also hold in the case where distributions over the positive and negative examples are distinguished (e.g., Valiant 1984).[8] While one crucial assumption necessary for PAC learnability is that of a finite VC dimension, there is a way to relax this assumption by allowing the sample size to be nonuniform for the different hypotheses placing unequal weight over the hypothesis classes. Technical details can be found in Section F of the Appendix.

4.2. Estimating VC dimension

The bound given by Equation 5 depends on the VC dimension of the target concept, which can generally be calculated analytically only for the most straightforward classifiers. To address this challenge, we calculate the sample complexity bound using an

[8] See Kearns (1990) for further discussion and proofs.

estimate that is consistent in simulation parameters rather than the true VC dimension. To do so, we estimate it empirically based on the known relationship between the worst-case generalization error of a classifier and its VC dimension, following McDonald, Shalizi, and Schervish (2011). Formally, the VC dimension of a hypothesis space $\mathcal{H}$ indicates the cardinality of the largest set $\mathcal{S}$ that can be *shattered*[9] by $\mathcal{H}$. Further details can be found in Appendix D.

In Figure 2, we present the results of this estimation procedure for a k-dimensional linear discriminant classifier, which is known to have a VC dimension of $k + 1$. The y-axis gives an estimated bound on the relationship between empirical risk and sample size for the given classifier, while the x-axis gives the sample size. Since the functional form of this relationship is known up to a constant given the true VC dimension, we can then estimate the VC dimension of any classifier through non-linear regression (Vapnik, Levin, and Le Cun 1994). McDonald, Shalizi, and Schervish (2011) demonstrates that this estimate is consistent in the number of simulations so that the estimate converges to the true VC dimension given sufficient computational resources.

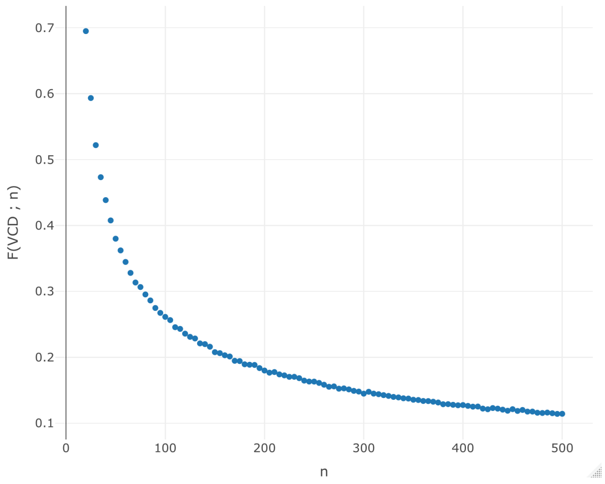


FIGURE 2. Simulation Study for Estimating the VC dimension of a linear discriminant model, $k = 7$

[9] "Shattering" is a key notion in machine learning that refers to a classifier's capacity to accurately distinguish any arbitrary labeling of a group of points.

To return to the example of polyarchies, the key idea is that the worst-case performance of any learning algorithm is when *all* of the examples available to it are wrongly labeled. Hence, it is sufficient to generate training data according to any data-generating process and then flip the labels. In this case, we would randomly assign a square region in the feature space to correspond to the classification boundaries of polyarchies and non-polyarchies, and then draw examples from it with the labels swapped. A classifier trained on this data then approximates the worst-case generalization error, since it is learning the opposite of what is needed. It is then sufficient to repeat this procedure for samples of increasing size in order to recover the classifier's VC dimension.

5. Simulation Studies

In order to verify that the sample complexity bounds perform reasonably well for realistic configurations of the parameters, we provide simulation-based analysis similar to Fong and Tyler (2021) and compare the empirical performance with theoretical bounds given the following inputs: (1) a confidence parameter (δ), (2) the accuracy parameter (ϵ) and (3) the misclassification rate (η). The simulations are generated by the following process:

Algorithm Simulation-based Analysis

```
1. Decide on desired accuracy parameters and concept
2. Calculate the VCD of the chosen model using the above estimation procedure
3. Generate a fine grid of points over the $k$-dimensional feature space
4. Classify these points according to the pre-defined concept
5. Generate observed labels by adding fixed independent random noise with probability $\eta$
6. Calculate sample complexity bounds empirically for a range of acceptable error rates
7. Repeat the process according to a range of values of "optimism" parameter (analytic bound corresponds to worst-case sampling)
```

We revisit the example of classifying polyarchy and apply our proposed approach as given by the Algorithm above. Values are calculated by fixing $\eta = 0.05$ and either $\epsilon = 0.06$ or $\delta = 0.01$. In this setting, Bound 5 gives a theoretical bound of 178 observations, which is borne out by the simulation. Notably, this bound is significantly higher than the sample size needed to achieve performance of over 95% on conventional out-of-sample performance metrics, since it involves the more stringent requirement that the probability of high error rates in *any* sample be controlled, and not simply the average misclassification rate. Although this may lead to more conservative conclusions regarding the target sample size, a key advantage is that researchers must explicitly specify the confidence δ with which they hope to achieve the desired out-of-sample accuracy, improving transparency and replicability. Morevoer, the theoretical bound of 178 cases closely corresponds to the simulation results under simple random sampling, shown in Figure 3, indicating that it is not unduly pessimistic even in this simple application.

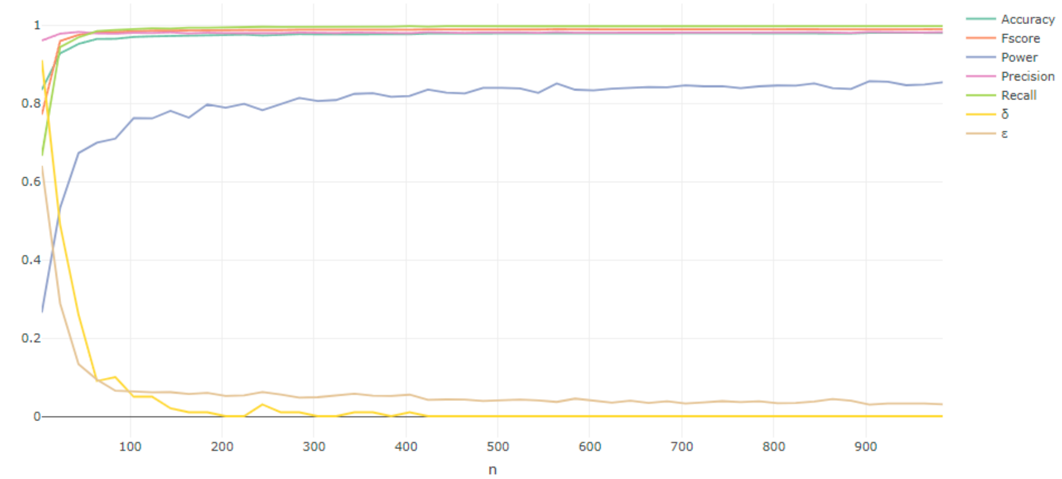


FIGURE 3. Learning to Classify Polyarchies

Figure 3 also shows the results of 1000 simulated experiments attempting to identify the effect of the latent concept – being a polyarchy – on an arbitrary outcome generated based on the true labels. Here, we hold sample size constant[10] and evaluate the ability

[10] Specifically, we hold the sample size of the test set, on which the experiment is conducted, constant.

of the trained model to correctly identify a significant effect ($\alpha = 0.05$) through out-of-sample predictions. As the training sample size increases and accuracy improves, the experimental power converges to the level (approximately 0.85) that would be achievable with 1000 observations if the latent concept were perfectly observed. Crucially, the results demonstrate that the training sample size needed to achieve the desired level of power in a downstream quasi-experiment again corresponds closely to the theoretical sample complexity bound for high confidence learning, further underlining the value of incorporating estimation error at the pre-analysis stage.

5.1. Limitations

Conventional classification tasks using machine learning have dealt with challenges associated with possibly incorrect functional form as well as estimation and measurement errors. However, the inference of our proposed methods is free from these possible concerns over the *a priori* beliefs on labeled data. In other words, our method does not make any strong functional form or distributional assumptions about true labels beyond the existence of i.i.d. measurement error and is therefore not subject to the usual inference challenges arising from potential sampling bias.

At the same time, our simulation approach is built on the assumption that the true data-generating process is perfectly learnable with the selected method given sufficient data without noise. In practice, much of the error in real-world applications comes not from measurement error, but from either misspecification error or irreducible noise. That is, the problem may not be a lack of data, but rather a fundamental mismatch between the target concept and the chosen algorithm, or even an inherent fuzziness in the concept itself that is not simply a problem of measurement error.

Since we assume away both of these very real issues, our method cannot speak directly to either of them, and the bounds we estimate may therefore be significant

The horizontal axis reflects variation in the size of the training set only.

underestimates when they are present. Worse, the pre-specified target accuracy may simply be unattainable in the presence of a large number of unmeasured variables or poorly selected models. The only advice we can offer to researchers in this regard is therefore to do their utmost to minimize the impact of such factors before performing our calculations, and to incorporate any residual uncertainty into their estimate of η.

Another important consideration is that our application of PAC learning focuses on binary-valued measures. While researchers can collapse continuous measures to binary by setting thresholds, a valuable extension would be to apply the framework to continuous and categorical concepts. Our analysis of learning can be readily extended to many other scenarios by allowing a variety of loss functions. For example, section I in the Appendix provides further intuition of sample complexity for tree-based models. The notion of sample complexity can undoubtedly accommodate various extensions. Next, although our validation exercises suggest that the theoretical bounds hold well in practice, one might still be concerned that the bounds may be too pessimistic. Related to this issue, one of our future tasks would be to update our simulation-based approach to allow researchers to parametrically pre-specify their "optimism" regarding the sampling procedure, generating an interval of progressively looser bounds.

6. Application to Predicting Recidivism (Dressel and Farid 2018)

We additionally offer an application of the method to actual noisy data, based on the study implemented by Dressel and Farid (2018). The original authors studied how predictions made by people with little or no criminal justice expertise can be comparable to machine-based commercial risk assessment software. The application of our method allows us to precisely assess the additional benefit provided by big data in this context, demonstrating how it could be applied by researchers prior to data collection. Our findings highlight the concept formation problem, as big data provides minimal additional benefit due to imprecise specification of the target concept.

Dressel and Farid (2018) compare the overall accuracy and bias in human assessment with the algorithmic assessment of COMPAS, a widely used algorithm for predicting recidivism. The authors hired 20 human coders recruited through Amazon's Mechanical Turk and used seven features (e.g., age, sex, number of juvenile misdemeanors, number of juvenile felonies, number of prior crimes, crime degree, and crime charge) for the analysis. We access the dataset used by Bansak (2019) to replicate Dressel and Farid (2018) and implement linear discriminant analysis as in the original paper. The model is trained on a random 80% of training and 20% testing split, with a VC dimension of 8. Note that in this case, we do not need to simulate the VCD dimension since it is known theoretically; however, it is straightforward to verify that our procedure produces the same value with a sufficiently large number of simulations.

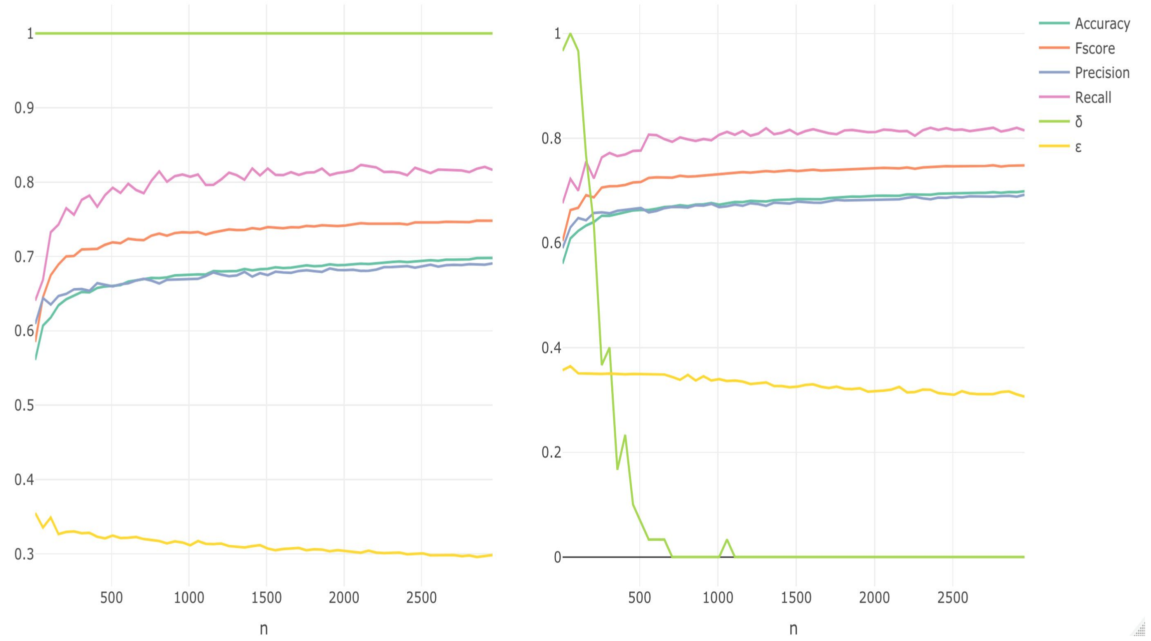


FIGURE 4. Simulation Analysis When $\epsilon = 0.05$ (Left) & $\epsilon = 0.35$ (Right)

Figure 4 shows the results of simulations on the observed data using $\delta, \eta = 0.05$, and values of 0.05 (left panel) and 0.35 (right panel) for ϵ. Note that in this case η, or the prevalence of labeling error, essentially corresponds to the false conviction rate since the data are produced directly from the criminal justice system. We use a value of 5%, since available evidence suggests a false conviction rate of between four and six percent (Gross et al. 2014).

The theoretical bound gives a target sample size of 272 observations for these parameters, which is again born out by the simulation results in the left panel. In practice, however, the results show that the best achievable accuracy with high confidence is approximately 35%, but the additional benefit of sample size above 500 is minimal. This application therefore highlights the central role played by the concept formation problem: the advantages of big data depend on precise specification of the target concept and selection of a corresponding learning algorithm.

As such, this application clearly demonstrates the advantage of the proposed method when used *prior to* data collection. While the original paper employs a dataset of almost 8000 observations – quite costly for many social science applications – Figure 4 demonstrates that the additional value of this large sample size is negligible given the methods employed. Although an alternative learning algorithm better suited to the target concept may have been able to take advantage of the additional observations, the combination of conceptual mismatch and algorithm choice ensured that most of the sample was essentially wasted. As with power analysis, the calculation of sample complexity bounds prior to undertaking research would prevent such situations from arising, ensuring that scarce research resources are allocated appropriately. We include supplementary exercises in Appendix G using data from Lewis et al. (2019). Since the second exercise yields results that are not significantly different from the key findings in Dressel and Farid (2018), we opt not to address it in the main text to economize space.

7. A Guide for Practitioners

As previously discussed, *a priori* analysis is used to calculate the necessary sample size N for PAC learning, given the error parameter ϵ, confidence parameter δ, and noisy rate η. While we emphasize the advantage of using the proposed methods prior to data collection, we do not exclude the possibility that researchers might select the "design" of sample complexity bounds, as shown in Table 2.

TABLE 2. Designs for Applying Sample Complexity Bounds

Type of Analysis	Input Parameters	Output
1. A priori	Accuracy (ϵ), Confidence (δ), Noisy rate (η)	N
2. Post-hoc	Confidence, Noisy rate, N	Accuracy

In addition to a priori analysis, sample complexity can bring benefits to researchers even for post-hoc analysis. While this approach is relatively less ideal as it controls only the confidence parameter rather than the accuracy parameter, it can still provide valuable insights into the accuracy of the measures produced by researchers using the given training sets. For instance, post-hoc sample complexity analysis can help determine whether the collected data was sufficient to support reliable inferences or if additional data is needed to achieve a given accuracy level with desired confidence out of sample.

As more researchers design and construct their own measures, understanding sample complexity can help them identify which quantities are more sensitive to researcher-specific parameter choices. We provide practical advice to practitioners below:

1. **Clearly define your concept:** The most important assumption in our framework is that there is a perfect match between the target concept and the classification algorithm employed by the researcher in the sense that the true classification rule lies within the hypothesis class $\mathcal{H}$ over which the algorithm is searching

and that the two *share the same VC dimension*. In addition, the bounds presented in this paper require that the concept be PAC-learnable to begin with, which is equivalent to assuming that it has a finite VC dimension. We provide guidelines for common cases where this assumption is violated – such as tree-based models and countable unions of finite classes – in the Appendix and incorporate them into the accompanying R package.

While this assumption is fundamentally untestable, it requires careful theoretical justification prior to algorithm selection in order to generate useful results. In other words, a given classification method should be selected not simply because it has been shown to be accurate in other research contexts, but because it is thought to be a good fit for the concept at hand. In particular, this requires careful theorization of concepts and reference to previous empirical work in order to ensure that the problem is well-defined: researchers should know *how many independent dimensions* there are to a concept, and how these tend to interact. For instance, in our running polyarchy example, a Dahlian view of regime type suggests that it can be well-represented as a two-dimensional linear interaction. It is precisely this kind of careful theorization that will yield precise results when assessing sample complexity.

2. **Justify your choice of research-specific parameters:** As in the choice of effect size for power analysis (Correll et al. 2020), we encourage practitioners to justify their choice of parameters using their domain expertise or drawing on existing studies in the past or empirical evaluation of the pilot data. This is particularly true of the misclassification rate, η, which has a significant impact on sample complexity. For instance, researchers studying democracy might begin by comparing the disagreement rates across existing measures in order to arrive at an approximate estimate of the rate of labeling error.

3. **Run simulation plots for pilot studies:** While drawing in domain knowledge and prior research can help to select appropriate parameter values, we reiterate the importance of conducting pilot studies where feasible. Using our companion R package,`scR`, researchers can run simulation analysis using their own pilot data. By visualizing the relationship between error rate and sample size, practitioners will be better positioned to make informed decisions on how to allocate resources.

If, after taking these steps, the resultant sample complexity of a classification task falls within the parameters of a proposed data collection effort, the analysis then lends weight to a research plan. This is particularly valuable to researchers at the grant application stage, as it can help to assure funders of its robustness. Conversely, when the analysis indicates that the sample complexity exceeds the budget or data availability, it provides an early warning to researchers prior to making costly investments. Such a discovery that the data is likely not to be *good enough* need not mean an injunction to abandon the project, however. Instead, the following courses of action are generally available to practitioners in this scenario. While the preferred choice dependent on the particular research context, we present them in order of decreasing order of desirability:

1. Researchers can simply gather more data, for example by applying for additional funding or locating new data sources. While this allows for pursuing the original research plan without compromise, it is frequently not practical due to scarce resources.

2. Researchers can adjust their target accuracy or confidence parameters or invest in reducing labeling errors in order to reduce the sample complexity to an acceptable level. In practice, this is essentially what widely-used post hoc approaches entail: estimation is optimized based on the best (estimated) out-of-sample error rate, which may fall short of the *ex-ante* target. However, there is a considerable advan-

tage to performing this step at the design stage, as it allows adjusting expectations based on what level of accuracy is likely to be achievable.

3. The third option is to revisit the choice of algorithm. As noted above, a key assumption underlying our method is the match between concept and classification algorithm. In practice, however, this is often not the case: many common classification tasks in political science are quite low-dimensional (Morucci and Spirling 2024), leading to a risk of selecting algorithms that are unnecessarily complex in terms of VC dimension. In other words, while the idea of applying a multi-billion parameter neural network capable of adjusting to even the most subtle concepts may be appealing, it is unnecessary if the true relationship can be well-represented by a linear discriminant and will result in unrealistically high sample complexity.

Perhaps the most important decision facing applied researchers involves the selection of confidence and accuracy parameters, ϵ and δ. As is evident from the simulation studies we provide (see, for example, Figure 3), reducing the confidence parameter δ typically imposes significantly more stringent requirements on the data than does ϵ. To see why this is the case, suppose that the target accuracy is 95%, corresponding to $\epsilon = 0.05$. Then, even as the *average* accuracy approaches this level, datasets will still be drawn with a high probability that result in marginally lower accuracy, say 94%. Setting δ close to 0 then requires that we control the probability of this occurring, which may result in hitting the point of diminishing returns with regards to *average* accuracy. For this reason, we would generally recommend that researchers choose a higher δ than ϵ; for example, we find $\delta = 0.2$, $\epsilon = 0.1$, or an 80% confidence of 90% accuracy to strike a reasonable balance in many cases.

Nevertheless, we strongly discourage researchers from simply applying these default values without further consideration. In practice, the appropriate values will largely be determined by the ultimate goal of the classification task. If the goal is purely descriptive

– for instance, to estimate the number of polyarchies in the world, or the proportion of progressive Democrats – then achieving high accuracy and confidence may not be particularly important. Conversely, when the classification has policy consequences, as in our recidivism outcome, tight control over error rates is desirable.

In addition, there are two important cases in which our method overlaps with power analysis for causal inference. The first of these is when the researcher seeks to identify the impact of a randomized treatment on a latent outcome; for instance, the impact of campaign donations on a politician's spoken ideology. The second instance arises in observational settings when the outcome is observed but the causal variable is latent, as in the impact of exposure to media ideology on survey respondents' voting intentions. In both of these cases, the accuracy with which the latent construct can be measured has direct consequences for the power of the causal estimator, since misclassification introduces noise, even if it is completely random. In these cases, we would therefore recommend a two-stage approach: first, researchers should identify the impact of classification error on the power of their causal estimator. Second, they should conduct a sample complexity analysis with accuracy and confidence parameters determined by the first stage in order to identify the additional demands on sample size introduced by the latent classification step.

8. Discussion

In this paper, we consider the question of what constitutes "good enough" data, both in terms of sample size and labeling accuracy, for statistical inference. To address this persistent question in applied research, we propose a novel application of PAC learning. We present the theoretical sample complexity bound that makes PAC learning feasible and provide a simulation-based approach to demonstrate its applicability, incorporating the VC dimension. The proposed method is a general-purpose tool for quantitative research in political science that allows researchers to make rigorous predictions about what can be "learned" from data before investing in costly data-collection efforts.

A key advantage of our approach is that it provides a more precise alternative to the assumption that the sample size is “large enough” for asymptotic approximations to hold. Furthermore, we directly consider the role played by labeling error and concept definition on model performance, a factor that has generally been overlooked in applied work, with hand-labeled data assumed to represent ground truth. Our approach thus provides a practical toolkit for researchers to guarantee adequate performance at the design stage, improving the replicability and external validity of studies reliant on machine learning classifiers.

Relatedly, we expect that sample complexity bounds will increase efficiency for researchers in terms of time and expenses. Traditional measurement approaches, such as hand-coding, tend to incur high startup costs (e.g., Quinn et al. 2010). Without tools to approximate the necessary sample size, researchers have often made decisions based on intuition within the scope allowed by their time and available resources. We hope that the application of sample complexity will add scientific rigor by providing statistical justification for the choice of sample size when applying machine learning to measurement tasks. While these bounds may not always be tight, they currently represent the best tool available to empirical researchers at the data collection stage. Although it is certainly possible for classifiers trained on fewer data to perform well, the aim of our paper is to provide theoretically informed guidelines to optimize the allocation of scarce resources.

Our goal is not to provide the definitive answer to the methodological question of what constitutes “good enough" data. Instead, we aim to initiate a dialogue. Social scientists have made significant strides in developing rigorous statistical models. However, the need to assess the quality of input has often been overlooked when performing downstream statistical inferences. We hope this paper contributes to rekindling attention to the importance of both data quantity and quality in applying statistical learning to social sciences.

9. Conflict of Interest

The authors are not aware of any conflicts of interest.

References

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo, and James A Robinson. 2019. Democracy does cause growth. *Journal of political economy* 127 (1): 47–100.

Arnold, Christian, Luka Biedebach, Andreas Küpfer, and Marcel Neunhoeffer. 2023. The role of hyperparameters in machine learning models and how to tune them. *Political Science Research and Methods,* 1–8.

Aslam, Javed A, and Scott E Decatur. 1996. On the sample complexity of noise-tolerant learning. *Information Processing Letters* 57 (4): 189–195.

Baltz, Samuel, Fabricio Vasselai, and Allen Hicken. 2022. An unexpected consensus among diverse ways to measure democracy. *Democratization* 29 (5): 814–837.

Bansak, Kirk. 2019. Can nonexperts really emulate statistical learning methods? a comment on "the accuracy, fairness, and limits of predicting recidivism". *Political Analysis* 27 (3): 370–380.

Barberá, Pablo, Amber E Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. Automated text classification of news articles: a practical guide. *Political Analysis* 29 (1): 19–42.

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review* 110 (2): 278–295.

Blumer, Anselm, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. 1989. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)* 36 (4): 929–965.

Carlson, David, and Jacob M Montgomery. 2017. A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review* 111 (4): 835–843.

Collier, David, and James E Mahon. 1993. Conceptual "stretching" revisited: adapting categories in comparative analysis. *American Political Science Review* 87 (4): 845–855.

Correll, Joshua, Christopher Mellinger, Gary H McClelland, and Charles M Judd. 2020. Avoid cohen's 'small','medium', and 'large'for power analysis. *Trends in cognitive sciences* 24 (3): 200–207.

Cox, Gary W, Jon H Fiva, and Daniel M Smith. 2020. Measuring the competitiveness of elections. *Political Analysis* 28 (2): 168–185.

Dahl, Robert A. 2008. *Polyarchy: participation and opposition.* Yale university press.

Dressel, Julia, and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4 (1): eaao5580.

Egami, Naoki, Musashi Hinck, Brandon Stewart, and Hanying Wei. 2024. Using imperfect surrogates for downstream inference: design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems* 36.

Esarey, Justin, and Andrew Pierce. 2012. Assessing fit quality and testing for misspecification in binary-dependent variable models. *Political Analysis* 20 (4): 480–500.

Fariss, Christopher J, and Zachary M Jones. 2018. Enhancing validity in observational settings when replication is not possible. *Political Science Research and Methods* 6 (2): 365–380.

Fong, Christian, and Matthew Tyler. 2021. Machine learning predictions as regression covariates. *Political Analysis* 29 (4): 467–484.

Grimmer, Justin. 2015. We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics* 48 (1): 80–83.

Grimmer, Justin, and Brandon M Stewart. 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* 21 (3): 267–297.

Grimmer, Justin, Sean J Westwood, and Solomon Messing. 2014. *The impression of influence: legislator communication, representation, and democratic accountability.* Princeton University Press.

Gross, Samuel R, Barbara O'brien, Chen Hu, and Edward H Kennedy. 2014. Rate of false conviction of criminal defendants who are sentenced to death. *Proceedings of the National Academy of Sciences* 111 (20): 7230–7235.

Hainmueller, Jens, and Chad Hazlett. 2014. Kernel regularized least squares: reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis* 22 (2): 143–168.

Hanneke, Steve. 2016. The optimal sample complexity of pac learning. *Journal of Machine Learning Research* 17 (38): 1–15.

Hollyer, James R, B Peter Rosendorff, and James Raymond Vreeland. 2014. Measuring transparency. *Political analysis* 22 (4): 413–434.

Hopkins, Daniel J, and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54 (1): 229–247.

Jerzak, Connor T, Gary King, and Anton Strezhnev. 2023. An improved method of automated nonparametric content analysis for social science. *Political Analysis* 31 (1): 42–58.

Kayser, Mark Andreas, and René Lindstädt. 2015. A cross-national measure of electoral competitiveness. *Political Analysis* 23 (2): 242–253.

Kearns, Michael J. 1990. *The computational complexity of machine learning.* MIT press.

Knox, Dean, Christopher Lucas, and Wendy K Tam Cho. 2022. Testing causal theories with learned proxies. *Annual Review of Political Science* 25 (1): 419–441.

Knuth, Donald E. 1976. Big omicron and big omega and big theta. *ACM Sigact News* 8 (2): 18–24.

Laird, Philip D. 2012. *Learning from good and bad data.* Vol. 47. Springer Science & Business Media.

Lesniewski, Niels, and Ryan Kelly. 2024. House gop had lowest win rate on "party unity" votes since 1982. February. https://rollcall.com/2024/02/08/house-gop-had-lowest-win-rate-on-party-unity-votes-since-1982/.

Lewis, Jeffrey, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. 2019. Voteview: congressional roll-call votes database. *See https://voteview. com/(accessed 27 July 2018).*

———. 2024. Voteview: congressional roll-call votes database. *https://voteview.com/.*

Little, Andrew, and Anne Meng. 2023. Subjective and objective measurement of democratic backsliding. *Available at SSRN 4327307.*

McDonald, Daniel J, Cosma Rohilla Shalizi, and Mark Schervish. 2011. Estimated vc dimension for risk bounds. *arXiv preprint arXiv:1111.3404.*

Miller, Blake, Fridolin Linder, and Walter R Mebane. 2018. Active learning approaches for labeling text. *Political Analysis.*

Morucci, Marco, and Arthur Spirling. 2024. Model complexity for supervised learning: why simple models almost always work best, and why it matters for applied research. *Working Paper.*

Neunhoeffer, Marcel, and Sebastian Sternberg. 2019. How cross-validation can go wrong and what to do about it. *Political Analysis* 27 (1): 101–106.

Poole, Keith T. 2005. *Spatial models of parliamentary voting.* Cambridge University Press.

Poole, Keith T, and R Steven Daniels. 1985. Ideology, party, and voting in the us congress, 1959–1980. *American Political Science Review* 79 (2): 373–399.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54 (1): 209–228.

Shalev-Shwartz, Shai, and Shai Ben-David. 2014. *Understanding machine learning: from theory to algorithms.* Cambridge university press.

Simon, Hans Ulrich. 1993. General bounds on the number of examples needed for learning probabilistic concepts. In *Proceedings of the sixth annual conference on computational learning theory,* 402–411.

Tian, Tian, and Jun Zhu. 2015. Max-margin majority voting for learning from crowds. *Advances in neural information processing systems* 28.

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

Vapnik, Vladimir, Esther Levin, and Yann Le Cun. 1994. Measuring the vc-dimension of a learning machine. *Neural computation* 6 (5): 851–876.

Vapnik, Vladimir N, and A Ya Chervonenkis. 2015. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis,* 11–30. Springer.

Vidyasagar, Mathukumalli. 2013. *Learning and generalisation: with applications to neural networks.* Springer Science & Business Media.

Wang, Siruo, Tyler H McCormick, and Jeffrey T Leek. 2020. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences* 117 (48): 30266–30275.

Ying, Luwei, Jacob M Montgomery, and Brandon M Stewart. 2022. Topics, concepts, and measurement: a crowdsourced procedure for validating topics as measures. *Political Analysis* 30 (4): 570–589.

Supporting Information for

Learning from Noise: Applying Sample Complexity for Political Science Research

Perry Carter[*] Dahyun Choi[†]

July 2024

[*] Ph.D. Candidate, Department of Politics, Princeton University.Email: pjcarter@princeton.edu

[†] Ph.D. Candidate, Department of Politics, Princeton University. Email: dahyunc@princeton.edu

A. When Learning is Noise-Free

While measurement models are typically optimized for accuracy on observed labeled instances, the ability to make *generalizable* claims beyond the training data is the ultimate goal of statistical learning. Although sample splitting is now widely used to address concerns of overfitting, it cannot substitute for having a sufficiently large dataset to begin with.

We formalize the problem as follows (Laird 2012): denote domain $\mathcal{X}$, label sets $\mathcal{Y}$, and concept classes $\mathcal{C}$. $\mathcal{X}$ includes all possible instances that the researcher may want to label and the set $\mathcal{Y}$ includes all possible labels or predictions for a single instance. An instance-label pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is called a labeled instance and a concept is a function $c : \mathcal{X} \rightarrow \mathcal{Y}$. We assume that there is an unknown concept[1] $\mathcal{C}$ that determines the true labels of instances.

Following the notation in the paper, we consider a probability distribution D over $\mathcal{X}$. We assume that the instances we observe are independent and identically distributed (i.i.d) according to an unknown D. Given that there is an unknown concept $\mathcal{C}$ which determines the true label of instances, the set of labeled instances $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ is generated by taking $x_i \sim D$ i.i.d and observing the corresponding $y_i = c(x_i)$. Then the true error and empirical error can be defined as follows. Suppose we have a model that produces a hypothesis $h \in \mathcal{H}$, given a sample of N training examples. The algorithm is called *consistent* if for every ϵ and δ, there exists a positive number of training examples $N \in \mathbb{N}$ such that for any distribution p^*, we have that

$$P(|R(h) - \hat{R}_N(h)| > \epsilon) < \delta \tag{6}$$

[1] A concept in this sense is precisely understood as a binary classification rule. This is not as divergent from common social science usage as it may initially appear: for instance, by the concept of "democracy", we mean a set of explicit rules that allow an observer to determine whether a given political system is or is not a *democracy*.

The *sample complexity* is the minimum value of N for which the equation (6) holds true. We therefore seek to determine the minimum sample size that produces a hypothesis within a specified error tolerance of the true concept with high probability. This is conceptually similar to the widely used approach of power analysis in experimental research and, as we show, has direct implications for power when researchers seek to identify the causal effect of a latent concept.

B. When Learning is Not Noise-Free

Here we provide further elaboration on Section 4.1 concerning the sample complexity bounds for applied research. This section provides a more detailed overview of the setup in Laird (2012) and Simon (1993) on which our approach is based. Suppose we have a rule e with error p. For example, the rule fails or disagrees with an example on average pm times in m examples. With the addition of noise, it may fail more often or less. Then the expected failure rate p_η is as follows:

$$p_n = (1-\eta)\,p + \eta(1-p) \tag{7}$$

The first indicates that if no classification errors occurs with probability $1-\eta$, the probability of failure is p. The second term suggests that if a classification error does not strike with probability η, the rule will fail only if it would not have failed without the error, with probability $1-p$. Note the cases below.

- When p =0 (zero error), its failure rate increases to η with noise.
- When $p \geq \epsilon$, its failure rate is at least $\eta + \epsilon(1-2\eta)$; and since $(1-2\eta) > 0$, this failure rate is greater than that of any correct one with zero error. [2]

We refer to rules with error greater than ϵ as ϵ-bad. Rules that are not ϵ-bad are ϵ-good. Rules with zero error are described simply as good. On average, ϵ-bad rules have a failure rate that is greater than that of good rules by at least $\epsilon(1-2\eta)$.

By the Law of Large Numbers, as $m \rightarrow \infty$

$$Pr[|p-\hat{p}| > \epsilon] \rightarrow 0$$

[2] When η is $\frac{1}{2}$, all of the information in the example is obliterated. When it is above one half, the algorithm would perform better by swapping all observed labels.

that is, $\hat{p}$ will be arbitrarily close to p for sufficiently many tests of an event whose probability of occurring is p, for all $\epsilon > 0$. If a correct hypothesis fails on average at the rate η, then with enough examples m, we will measure a failure rate closest to η with high probability. Similarly, an ϵ- bad rule will fail at nearly its expected rate, $\eta + s$, where $s \geq \epsilon(1 - 2\eta) > 0$.

LEMMA 1 (Hoeffding's inequality). *Consider a Bernoulli random variable with probability p of having the value 1 and $1 - p$ of having value 0. Let $GE(p, m, r)$ be the probability of at least $\lceil rm \rceil$ successes in m independent trials, and $LE(p, m, r)$ be the probability of at most $\lfloor rm \rfloor$ successes. If $0 \leq p \leq 1$, $0 \leq s \leq 1$, and m is any positive integer then*

$$LE(p, m, p - s) \leq e^{-2s^2 m}$$

and

$$GE(p, m, p + s) \leq e^{-2s^2 m}$$

This lemma bounds the probability that r, the empirical rate of success, is very different from p.

Suppose $\epsilon > 0, \delta \leq \frac{1}{2}$ and $0 \leq \eta \leq \eta_b < \frac{1}{2}$. Let success for a rule e_i refer to the event of disagreeing with a random sample. In m examples, F_i is the number of successes, and $\frac{F_i}{m}$ is the empirical rate of disagreement.

THEOREM 1. *When*

$$m \geq \frac{2}{\epsilon^2(1 - 2\eta_b)^2} ln \frac{2N}{\delta}$$

, the algorithm pac-identifies the concept $\mathcal{H}$.

Or a tighter bound:

THEOREM 2. *Let* $\eta < 1/2$ *be the rate of classification noise and N the number of rules in the class* $\mathcal{E}$. *Assume* $0 < \epsilon, \delta < \frac{1}{2}$. *Then the number m of examples required is at least*

$$m \geq \max \left[\frac{\ln(1/2\delta)}{\ln[1 - \epsilon(1 - 2\eta)]^{-1}}, \log_2 N(1 - 2\epsilon(1 - \delta) + 2\delta) \right] \tag{8}$$

and at most

$$m \leq \frac{\ln(N/\delta)}{\epsilon \left[1 - \exp[-\frac{1}{2}(1 - 2\eta)^2] \right]} \tag{9}$$

Proofs

The probability that a good rule fails on an example is η, while the probability that an ϵ-bad rule fails is at least $\eta + \epsilon(1 - \epsilon)$. The difference between these two rates is at least $\epsilon(1 - \eta) \geq \epsilon(1 - 2\eta_b) = s$. The algorithm is poor if some ϵ-bad rule happens to fail less often than all acceptable rules. Consider such an ϵ-bad rule, e_i, and let e_t be a correct rule. Let F_i and F_t be their respective failure statistics. By Hoeffding's inequality above, the probability of the first of these is at most,

$$\begin{aligned} LE(\eta + s, m, \eta + s - \frac{s}{2}) &\leq e^{-2(s/2)^2 m} \\ &\leq \frac{\delta}{2N} \end{aligned}$$

Similarly, the latter is

$$GE(\eta, m, \eta + \frac{s}{2}) \leq \frac{\delta}{2N}$$

Thus, the probability that any ϵ-bad rule e_i fails less than e_t is at most $\frac{\delta}{N}$. Since there are fewer than N bad rules, the probability that one of them minimizes the number of failures by the algorithm is less than δ. The theorem below addresses a more generalizable learning model when instances are not noise-free. For further details, see Simon (1993)'s formalized proofs, particularly Corollary 3.13.

THEOREM 3. *Let C be a class of concepts and $VC(C) \geq 2$ its Vapnik-Chveronenkis dimension. Any algorithm that learns C with regard to classification noise rate η needs $\Omega(\frac{VC(C)}{\epsilon(1-2\eta)^2})$ observations.*

Proof The sample for f is defined as the sample of a p-concept f_η, where $f_\eta(x) = \eta$ if f(x)=0, and $f_\eta(x) = 1 - \eta$ otherwise. Let $C_\eta = \{f_\eta | f \in C\}$. Let $d = d(C)$ and S be the sequence of size d which is shattered by $\mathcal{C}$. S is γ-shattered by C_η for $2\gamma = 1 - 2\eta$. Let A be an algorithm which learns $\mathcal{C}$ under classification noise rate η. The domain distribution D can be chosen as in the proof of Theorem 3.1. If A's output $\bar{h}$ is a hypothesis for target concept f whose error is bounded by ϵ, then $\mathcal{H}$ is an $(\epsilon, 0)$-good model for f_η on S.

C. Uniform Convergence and ϵ- representative sample

LEMMA C1. *Assume that a training set S is $\frac{\epsilon}{2}$-representative for the domain, loss function, and distribution D. Then, any hypothesis $h_s \in argmin_{h\in\mathcal{H}} : L_s(h)$ satisfies the following condition.*

$$R(h_S) \leq min_{h\in\mathcal{H}}\hat{R}(h)| + \epsilon \tag{10}$$

The proof is as follows. For every $h \in \mathcal{H}$,

$$R(h_S) \leq \hat{R}(h_s) + \frac{\epsilon}{2} \leq \hat{R}(h) + \frac{\epsilon}{2} \leq R(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = R(h) + \epsilon \tag{11}$$

The second equality holds because h_s is a predictor that reduces empirical risk. And the assumption that the S is $\frac{\epsilon}{2}$ representatives ensures that the third and first equalities hold.

The formal definition of the uniform convergence property is as follows. A hypothesis class $\mathcal{H}$ has the uniform convergence property regarding its domain and loss function if there exists a function $m_H : (0,1)^2 \rightarrow N$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution D over domain $\mathcal{X}$, if S is a sample of $m > m_{\mathcal{H}}$ examples drawn according to D, then with probability of at least $1 - \delta$, S is ϵ- representative. The function $m_{\mathcal{H}}$ is equivalent to the sample complexity, the minimal necessary sample size of obtaining the uniform convergence property. Again, the term *uniform* refers to having a fixed sample size that works for all members of H and over all possible probability distributions over the domain.

D. Estimated VC dimension for risk bounds

McDonald, Shalizi, and Schervish (2011) propose a simulation-based method to estimate the VC dimension, which measures the generalization capacity of learning algorithms. They prove two main results: first, that the estimated VC dimension will concentrate around the true dimension with high probability, and second, that using the estimated VC dimension allows for the recovery of accurate bounds on generalization error.

Building on Vapnik, Levin, and Le Cun (1994), which shows that the expected maximum deviation between the empirical risks of a classifier on two datasets can be bounded by a function that depends only on the VC dimension of the classifier, the authors provide the following function of n and parametrized by $\mathcal{H}$:

$$VC(F) = \begin{cases} 1 & n < \frac{h}{2} \\ a\frac{l\,og\frac{2n}{h}+1}{\frac{h}{h}-a''}\left(\sqrt{1+\frac{a'(\frac{n}{h}-a'')}{l\,og\frac{2n}{h}+1}}+1\right) & \text{else} \end{cases}$$

Following Vapnik, Levin, and Le Cun (1994) ,the constants were chosen as follows: a = 0.16, $a^{'}$ = 1.2 and $a^{''}$ = 0.14927 so that ϕ (.5) = 1. These values are tuned to be optimal for linear discriminant classifiers, for which the VCD is known theoretically. A key assumption is therefore that the same values can be used for the chosen algorithm without introducing bias.

As we have imperfect knowledge, we generate many observations

$$\hat{\xi}(n) = \Phi_h(n) + \epsilon(n)$$

along a fine grid of design points n. Here ϵ is centered on mean zero as the bound is tight, having an unknown distribution with support on [0,1]. We then estimate the true

VC dimension h^* using nonlinear least squares. Of course, generating $\hat{\xi}(n_l)$ is nontrivial. Vapnik, Levin, and Le Cun (1994) provides an algorithm for generating the appropriate observations. At each fixed design point $n_l : l \in \{1, ..., k\}$, we simulate m data points for i = 1,....,m, so as to approximate $\xi(n_l)$ as defined. Vapnik, Levin, and Le Cun (1994) shows that this approach works well in practice, recovering the known VC dimension of linear classifiers and demonstrates that the method for generating the dataset does not affect the algorithm's performance, since for any data structure it is sufficient to flip labels to ensure the most inaccurate possible algorithm is trained.

Below is the procedure for generating $\hat{\xi}(n_l)$, discussed in Vapnik, Levin, and Le Cun (1994), which we apply here.

Algorithm generating $\hat{\xi}(n_l)$

```
Given a collection of possible classifier F and a grid of design points, repeat the procedure at each design point, $n_l$, m times

1. Generate a dataset from the same space $y \times X$ as the training sample that is independent of the training sample. The generated set should be of size $2n_l$.
2. Split the data set into two equal sets W and $W^{'}$
3. Flip the labels (y values) of $W^{'}$
4. Merge the two sets and train the classifier simultaneously on the entire set: W with the “correct" labels and $W^{'}$ with the “wrong" labels.
5. Calculate the training error of the estimated classifier $\hat{f}$ on W with the correct labels and on $W^{'}$ with the wrong labels.
6. Set $\hat{\xi}(n_l) = |\hat{R}_{n_l}(\hat{f}, W) - \hat{R}_{n_l}(\hat{f}, W^{'})|$.
7. Set $\hat{\xi}(n_l) = \frac{1}{m}\sum_{i=1}^{m} \hat{\xi}(n_l)$
```

E. Proofs of Distribution-free Property and Learnability

This section presents proofs that establish the necessary and sufficient conditions underpinning the distribution-free assumption of the PAC model (Vapnik, Levin, and Le Cun 1994; Vapnik and Chervonenkis 2015). We use the mathematical notations in learning theory and proofs as in **?**.

We define the concept class as a nonempty set $\mathcal{C} \subseteq 2^X$. X is a fixed set, either finite or countably infinite, of binary indicators which can be mapped into Euclidean n-dimensional space for $n \geq 1$. We also assume that each $c \in C$ is a Borel set. The sample space of $\mathcal{C}$, denoted by $S_{\mathcal{C}}$ is the set of all m-samples over all $c \in C$ and all $x \in X^m$ for all $m \geq 1$.

$A_{\mathcal{C},\mathcal{H}}$ indicates the set of all functions A that maps $S_{\mathcal{C}}$ into $\mathcal{H}$, where $\mathcal{H}$ is a set of Borel Sets on X. $\mathcal{H}$ is the hypothesis space. Elements in $\mathcal{H}$ are named hypotheses. We allow A to approximate concepts in $\mathcal{C}$ using hypotheses from a different class $\mathcal{H}$. A is consistent if its hypothesis always agrees with the sample, whenever $h = A$ then for all i, $1 \leq i \leq m$, $i = I_h(x_i)$. For any A, probability distribution P on X, $c \in C$, and $x \in X$, the error of A for concept $\mathcal{C}$ on $\tilde{x}$ is given by $P(c \triangle h)$

Definition 4. *Given a nonempty concepts class $\mathcal{C} \subseteq 2^X$ and a set of points $S \subseteq X$, Π denotes the set of all subsets of S that can be obtained by intersecting S with a concept in $\mathcal{C}$. If $\Pi = 2^s$, then we say S is shattered by $\mathcal{C}$. The VC dimension of $\mathcal{C}$ is the cardinality of the largest finite set of points $S \subseteq X$ that is shattered by $\mathcal{C}$. If arbitrarily large finite sets are shattered, the VC dimension of $\mathcal{C}$ is infinite.*

The empty set is always shattered, therefore, we also derive the following definition.

Definition 5. *For any integer $m \geq 0$, Π indicates the max over all $S \subseteq X$ of cardinality m.*

Based on this definition, the VC dimension of $\mathcal{C}$ can be defined as the largest integer

d such that $\Pi(d) = 2^d$ or infinity. We now present the summary of theories and provide the relevant proofs from Blumer et al. (1989).

THEOREM 4. *Let $\mathcal{C}$ be a concept class, then $\mathcal{C}$ is uniformly learnable if and only if the VC dimension of $\mathcal{C}$ is finite.*

Proof: Blumer et al. (1989) shows that the "if" part of the theorem follows as we always produce a consistent function that maps from $S_{\mathcal{C}}$ to $\mathcal{C}$ by simply well ordering the concept in $\mathcal{C}$ and choosing for each sample in $S_{\mathcal{C}}$ the first concept that is consistent with the sample. Likewise, the "only if" part of the theorem above holds as the lower bound grows arbitrarily large with d for the appropriate choice of ϵ and δ. Detailed proofs are provided on page 936 of Blumer et al. (1989).

F. Nonuniform learnability

We start with the definition of uniform convergence, as provided by Lemma C1. It naturally follows that for every sample complexity and δ, with a probability of at least $1 - \delta$, we have the following equation.

$$\forall h \in \mathcal{H}, |R(h) - \hat{R}_S(h)| < \epsilon(\delta) \tag{12}$$

Let w be a function such that $\sum_{n=1}^{\infty} w(n) \leq 1$. w indicates a weight function over the hypothesis classes and it reflects the importance of each hypothesis class or the relative complexity of different hypothesis classes. If $\mathcal{H}$ is a finite union of N hypotheses classes, researchers can simply assign the same weight of $\frac{1}{N}$ to all hypothesis classes. Equally weighting hypothesis classes is equivalent to empirical risk manipulation over a weighted superclass of hypothesis classes. If researchers have prior knowledge that a particular hypothesis is more likely to contain correct target concepts, then they should assign a larger weighting. However, uniform weighting does not work if $\mathcal{H}$ is an infinite union of hypothesis classes.

Therefore, researchers should rely on another framework to minimize the structural risks. This is a bound minimization approach, meaning that the goal of the framework is to figure out the hypothesis that minimizes a certain *upper* bound on the true risk. The bound that this framework hopes to minimize is provided in the following theorem.

Formally, let there be a weighting function, such that $\sum_{n=1}^{\infty} w(n) \leq 1$. Let $\mathcal{H}$ be a hypothesis class which is an infinite union of n hypotheses where each hypothesis satisfies the uniform convergence property with a sample complexity function. Let ϵ_n indicate the lowest possible upper bound on the gap between empirical and true risks achievable by using a sample of m examples.

$$\epsilon_n = \min\left\{\epsilon \in (0,1) : m(\epsilon, \delta) < m\right\} \tag{13}$$

Then, we can derive the following theorem.

THEOREM 5. *For every $\delta \in (0,1)$, and distribution D, with a probability of at least $1-\delta$, the following bound holds for every $n \in \mathcal{M}$ and $h \in \mathcal{H}$.*

$$|R(h) - \hat{R}_S(h)| < \epsilon(w(n), \delta)$$

Therefore, for every $\delta \in (0,1)$ and distribution *D*, with a probability of least $1-\delta$, the following equation holds.

$$\forall h \in \mathcal{H}, R(h) \leq \hat{R}_S(h) + \epsilon(w(n), \delta) \tag{14}$$

Proof: For each n, we define δ_n being equal to $w(n)\delta$. Relying on the assumption that uniform convergence holds for all n, we obtain that if we fix n in advance, then with a probability of at least $1 - \delta_n$ over researchers' choices. Then, applying the union bound over n = 1, 2, ..., the preceding holds for all n, with a probability of at least $1 - \sum_n \delta_n = 1 - \delta \sum_n w(n) \geq 1 - \delta$.

G. Application to Predicting Ideology (Lewis et al. 2024)

We further examine the validity of the proposed methods using the data provided by Lewis et al. (2024). Poole and Daniels (1985) and Poole (2005) developed basic data and measurements related to the United States Congress, using roll-call voting data. In this exercise, we examine the generalizable inference of roll-call votes on *NOMINATE* scores. To reduce simulation time, we focus on roll call votes since the 114th Congress.The model is trained with a VC dimension of 3. Since the VC dimension can be analytically calculated, we do not need to simulate the VCD.

Figure G.1 shows the results of the simulation on the observed data using δ, $\eta = 0.05$, and values of 0.01 (left panel) and 0.4 (right panel) for ϵ. In this case, η could correspond to rates of roll-call votes that might not match party ideology. While it could be as high as approximately 40% according to Lesniewski and Kelly (2024), we use 0.05 for δ for easy comparison with our first application from Dressel and Farid (2018). The simulation shows that the best achievable accuracy is 1 when ϵ is 0.01, and the additional benefits of a sample size above 500 appear trivial. This application also conveys similar insights from the previous application, indicating that the advantages of big data might diminish for generalization problems using simple models.

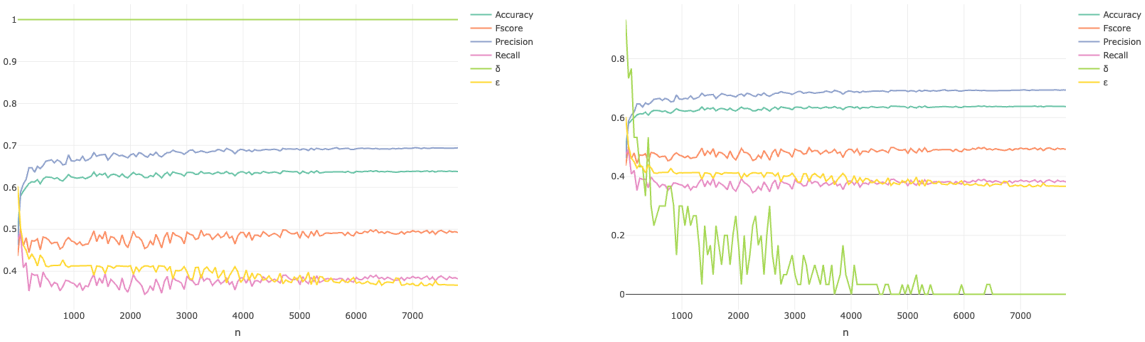


FIGURE G.1. Simulation Analysis When $\epsilon = 0.01$ (Left) & $\epsilon = 0.4$ (Right)

H. R Package: `scR`

To facilitate application, we provide an R package, `scR`, that provides a computationally efficient way to implement the proposed methods. *scR* makes use of an interface to the popular *caret* package in R to streamline the model training process for classification problems. If the VCD of the chosen algorithm is already known, users can use the `scb()` function directly to calculate sample complexity bounds given inputs ϵ, δ, η. Alternatively, users can interface with the rest of the package by defining their chosen algorithm any custom function that takes `formula` input and outputs a binary classification vector. This can then be fed to the `simVCD()` function to automatically calculate the VC dimension using the methods presented in this paper, which typically runs in under an hour. Alternatively, if the VC dimension is left unspecified, researchers can simply allow the `scb()` function to calculate bounds directly through simulation using the Algorithm above.

`scR` additionally supports a number of utility functions to generate simulation plots to see the relationship between researcher-specific parameters and the necessary sample size. To mitigate the concerns that the bound could be overly conservative, we allow researchers to explore multiple options for sampling regimes, ranging from i.i.d. to adversarial selection of examples far from the classification boundary.

I. Sample Complexity for Decision Trees

A decision tree predicts the label associated with an instance x by traveling from the root node of a tree to a leaf. While decision trees can be applied in a variety of settings, including multiclass problems, we focus on the binary classification setting, namely $y \in \{0, 1\}$, for simplicity. At each node on the root-to-leaf path, the successor is decided by the basis of splitting the input space using either one of the features or a predefined set of splitting rules.

In this case, we focus on one of the most widely used rules: thresholding the value of a single feature. We move between the right or left child of the node on the basis of $1_{\chi_i<\theta}$, where χ_i is an indicator for relevant features and $\theta \in R$ indicates the threshold. Then intuitively, a tree with k leaves can shatter a set of k instances. Therefore, if we assume decision trees of arbitrary size, we obtain a class of infinite VC dimension which can easily lead to overfitting and – most concerningly for our application - an infinite sample complexity bound.

To avoid this, we can rely on the minimum description length (MDL) principle where hypotheses with shorter descriptions are preferred, following the principle of Occam's razor. While this section does not go over the details of how MDL principle is constructed, the intuition is quite straightforward. We aim to learn a decision tree that fits the data while not being excessively complex.

We focus on the following scenario to explore how nonuniform learnability can be applied to tree models. For simplicity, we assume that $\chi \in \{0, 1\}^d$. In that case, thresholding the value of a single feature corresponds to a splitting rule. This notational simplification is without loss of generality and the following analysis can be applied to general cases.

With the simplifying assumption above, the classifier's dimension becomes finite, though it can still be very large. For example, any classifier from $\{0, 1\}^d$ to $\{0, 1\}$ can be represented by a decision tree with 2^d leaves and depth of $d + 1$. Therefore, the VCD becomes 2^d, which suggests that the number of examples we need to PAC learn the hypothesis class grows with 2^d. Unless d is very small, this leads to a sample size that is generally unattainable in social science.

To address this potential issue of overfitting, we rely on the MDL framework. Below is a theorem fromShalev-Shwartz and Ben-David 2014.

THEOREM 6. *Let $\mathcal{H}$ be a hypothesis class and let $d : \mathcal{H} \rightarrow 0, 1^*$ be a prefix-free description language for $\mathcal{H}$. Then, for every sample size, m, every confidence parameter, $\delta > 0$ and every probability distribution D, with probability greater than $1 - \delta$ over the choice of $S \sim D^m$ we have that,*

$$\text{For all } h \in \mathcal{H}, L_D(h) \leq L_s(h) + \sqrt{\frac{|h| + \ln(2/\delta)}{2m}}, \tag{15}$$

where $|h|$ is the length of d(h).

Intuitively, we prefer smaller trees over larger trees; to formalize this intuition, we first need to define a description language for decision trees, which is prefix-free and requires fewer bits for smaller decision trees. In principle, this could simply be plain English descriptions of the trees. One possible approach we can take that provides greater precision is as follows: A tree with n nodes will be described in n+1 blocks, each of size $l\,og_2(d + 3)$ bits. The first n blocks encode the nodes of the tree, in a depth-first order, and the last block marks the end of the code. Each block indicates whether the current node is:

- An internal node of the form $1_{x_i=1}$ for some $i \in [d]$
- A leaf whose value is 1
- A leaf whose value is 0

- End of the node

Overall, there are d+3 options, hence we need $log_2(d+3)$ bits to describe each block. Assuming each node has two children, without loss of generality,[3] we can show that this is a prefix-free encoding of the tree, and the length of a tree with n nodes is (n+1)log_2(d+3).

By Theorem 6, we have that, with a probability of at least $1 - \delta$ over a sample of size m, for every n and every decision tree h ∈ H with n nodes it holds that

$$L_D(h) - L_S(h) \leq \sqrt{\frac{(n+1)log_2(d+3) + log(2/\delta)}{2m}} \tag{16}$$

where the left-hand side again indicates the gap between true and empirical risks.

This bound reflects the key tradeoff: we expect more complex and larger decision trees to have a smaller training risk, $L_s(h)$, but the respective value of n will be larger. However, smaller decision trees will have a smaller value of n, but $L_s(h)$ might be larger. Researchers are hoping to find a decision tree with both low empirical risk, $L_s(h)$ and a number of nodes n not too high. The bounds suggests that such a tree will have low true risk, $L_D(h)$, allowing us to overcome the problem of non-finite VC dimension due to overfitting.

[3] If a decision node has only one child, we can still replace the node by its child without affecting the prediction of the decision tree.