A Rademacher Complexity Based Method for Controlling Power and Confidence Level in Adaptive Statistical Analysis

Lorenzo De Stefani
Department of Computer Science
Brown University
Providence, United States of America
lorenzo@cs.brown.edu

Eli Upfal
Department of Computer Science
Brown University
Providence, United States of America
eli@cs.brown.edu

Abstract—**While standard statistical inference techniques and machine learning generalization bounds assume that tests are run on data selected independently of the hypotheses, practical data analysis and machine learning are usually iterative and adaptive processes where the same holdout data is often used for testing a sequence of hypotheses (or models), which may each depend on the outcome of the previous tests on the same data. In this work, we present RADABOUND a rigorous, efficient and practical procedure for controlling the generalization error when using a holdout sample for multiple adaptive testing. Our solution is based on a new application of the Rademacher Complexity generalization bounds, adapted to dependent tests. We demonstrate the statistical power and practicality of our method through extensive simulations and comparisons to alternative approaches. In particular, we show that our rigorous solution is a substantially more powerful and efficient than the differential privacy based approach proposed in Dwork et al. [1]–[3].**

Index Terms—**Adaptive Analysis, Rademacher Complexity, Statistical Learning**

I. INTRODUCTION

The goal of data analysis and statistical learning is to model a stochastic process, or distribution, that explain an observed data. A major risk in statistical learning is *overfitting*, that is, learning a model that fits well with the observed data but does not predict new data. The standard practice in machine learning is to split the data into *training* and *holdout* (or testing) sets. A learning algorithm then learns a model using the training data and tests the model on the holdout set to obtain a confidence interval for the expected error or for the value of the loss function of the model. If the process halts after a single iteration, then the statistical analysis is easy. However, in most cases, the learning process is iterative and adaptive. One uses successive tests for model selection, feature selection, parameter tuning, etc., and the choice of the tests themselves often depends on the outcomes of previous tests. Ideally, each hypothesis should be tested on a fresh data sample. However, it is common practice to reuse the same holdout data to evaluate a sequence of hypotheses. While

widespread, this practice is known to lead to *overfitting*; that is, the learned model becomes representative of the sample rather than the actual process. Evaluating the accumulated error in testing a sequence of related hypothesis on the same data set is a major challenge in both machine learning and modern statistics. In machine learning, the problem of "*preventing overfit*", is usually phrased and analyzed in terms of bounding the generalization error [4]. In inference statistics, the goal is controlling the Family Wise Error Rate (FWER), or the False Discovery Rate (FDR) of a sequence of hypothesis tests [5].

Our Results: We develop and analyze RADABOUND, a rigorous, efficient, and practical procedure for online evaluation of the accumulated generalization error in a sequence of statistical inferences applied to the same sample. RADABOUND can evaluate fully adaptive sequences of tests. The choice of a test may depend on the information obtained from previous tests, and the total number of tests is not fixed in advance.

One way to quantify the risk of overfitting after k queries is by considering the probability of the condition defined by the results of the first k queries. If the probability of such condition is close to 1, the results of the queries evaluated so far do not significantly restrict the sample space, and there is, therefore, no risk of overfitting. Viceversa, if the probability of the observed condition is small, the sample space defined by the true distribution conditioned on the results of the queries is noticeably different from the true distribution, and there is thus a significant risk of overfitting. In general, it is hard to bound the probability of the observed condition as the true distribution over the samples is unknown. However, in the special case for which the queries being considered correspond to evaluating the average of functions (such as evaluating the average risk or loss functions of alternative learning procedures), we can design an adaptive process based on an empirical estimate of the Rademacher Complexity of the set of queries which correctly bounds this probability and correspondingly halts the procedure when the risk of overfitting exceeds a certain threshold fixed by the user.

IEEE
computer
society

Our method builds on the concept of Rademacher Complexity [6], [7] that has emerged as a powerful alternative to VC-dimension and related uniform convergence methods for characterizing generalization error and sample complexity. A fundamental advantage of the Rademacher Complexity approach in contrast to standard uniform convergence tools, such as VC-dimension, that capture the complexity with respect the worst case input distribution, is that it yields a data-dependent bound as it is computed with respect to the input (sample) distribution, and can be efficiently approximated from the sample.

Our solution employs three major components: (1) For a set of functions chosen independent of the sample, the Rademacher Complexity [6], [7] provides a powerful and efficient bound on the error in estimating the expectations of all these function using one sample; (2) As long as the outcome of the sequence of tests does not significantly overfit to the sample, conditioning on these outcomes has only a minor effect on the distribution; and (3) The Rademacher Complexity of a sequence of tests can be estimated efficiently form a given sample, requiring similar computation time as running the actual tests. To fully utilize our technique, we need computationally efficient methods for rigorously estimating the Rademacher Complexity from a sample. We introduce two novel methods based on Bernstein's inequality for martingales [8] and the Martingale Central Limit Theorem [9]. Our analysis and extensive experiments prove and demonstrate that our method guarantees statistical validity while retaining statistical power and practical efficiency.

Related Work: Classic statistics offers a variety of procedures for controlling the Family Wise Error Rate (FWER), ranging from the simple Bonferroni [10] to Holm's step-down [11] and Hochberg's step-up procedures [12] in the context of multiple hypotheses testing. While controlling the FWER under weak assumptions about the hypotheses, these methods are too conservative, giving many false negative results, in particular for large sets of hypotheses. Less conservative procedures, such as Benjamini and Hochberg [5], which control the False Discovery Rate (FDR) (i.e., the expected fraction of false discoveries), still do not scale up well for a very large number of hypotheses. However, all these procedures cannot be applied in the adaptive setting, as they require for the set of hypotheses to be fixed at the beginning of the testing procedure (i.e., before any data evaluation).

In statistics, "*sequential analysis*" or "*sequential hypothesis testing*" is a paradigm for statistical testing where for a fixed family of hypotheses to be the tested the sample size is not fixed in advance. Instead, data are evaluated as they are collected, and further sampling is stopped in accordance with a pre-defined stopping rule as soon as significant results are observed. Despite the sequential iterative nature of this practices, as the hypotheses being considered are fixed beforehand, sequential analysis procedures are not suitable for adaptive analysis as the set of queries (hypotheses) being considered depends in general for the outcome of previous evaluation of

the data itself. Other "*sequential*" hypothesis testing procedures, such as the *sequential False Discovery Rate* control by G'Sell et al. [13], assume that the order according to which the hypotheses are to be evaluated is fixed beforehand, and hence cannot be adaptively selected. Similar considerations apply to the "*Alpha Investing*" sequential testing by Foster and Stine [14], which achieves control of the "*marginal False Discovery Rate*". While the previously mentioned procedures apply to the setting of hypotheses testing, the method proposed in this work allows adaptive evaluations of statistical queries while maintaining rigorous guarantees on the accuracy of the obtained estimates.

A series of recent papers [1]–[3] explored an interesting relation between "*Differential Privacy*" [15] and overfit prevention in adaptive analysis. The basic idea is to limit the user access to the holdout data so that the answers to the sequence of queries is differentially private. A differentially private access to the holdout data limits the risk of overfitting to that data set. Unfortunately, the practical application of this elegant mathematics is limited. Differential privacy is achieved through random perturbation of the data (or the reply to the queries). The higher the number of adaptive queries, the larger the required perturbation. However, the amount of perturbation is limited by the need to preserve the actual signal in the data. As a result, rigorous application of this approach is either limited to a small number of queries or is computationally intractable [3], making it less useful than alternative methods [16], [17]. Our experiments in Section 5 show that RADABOUND allows orders of magnitude reduction of the required holdout dataset compared to Dwork et al.'s method [1] while offering the same guarantees. Further, our technique is much simpler as it does not require any introduction of additional noise. We discuss in detail the advantages of our solution compared to [1] in Section VI. A more practical solution for a restricted setting inspired by machine learning competitions was presented in [18]. Their solution, "*the Ladder*", provides a loss estimate only for those that made a significant improvement over the previous best. This restricted setting allows to sidestep the hardness results discussed in [16], [17]. Note however that the guarantee achieved by the Ladder is fundamentally different from the one achieved in Dwork et al. [2] and in this work, as it ensures accuracy in the relative ordering of the performance of multiple classifiers, while the latter ensure accurate evaluations of adaptively selected queries. Hence, the former is not comparable with the latters.

Paper organization: The presentation is organized as follows: In Section II we introduce out `RadaBound` method for adaptive statistical analysis. In Section III we discuss the use of uniform convergence bounds based on Rademacher Complexity in our setting, and we present two methods for estimating the Rademacher Complexity of a class of adaptively selected functions from the data. In Section IV we present the details of our `RadaBound` and the guarantees provided by it. In Section V, we present an experimental validation of the

correctness and power of our methods using synthetic data. Finally, in Section VI we compare our approach to the state-of-the-art approach based on *Differential Privacy* by Dwork et al. [2]. In particular, we show that a Rademacher complexity based solution gives significantly better results than the more complicated differential privacy based solution of Dwork et al..

II. THE RADABOUND

For concreteness, we focus on the following setup. We have an holdout sample composed by m independent observations $\bar{x} = (x_1, \ldots, x_m)$, each from a distribution $\mathcal{D}$, and parameters $\epsilon, \delta \in (0, 1)$ fixed by the user.

The process: In an iterative process, at each step, the user (or an adaptive algorithm) submits a function f and receives an estimate $\tilde{E}_{\bar{x}}[f] = \frac{1}{m} \sum_{i=1}^m f(x_i)$ of the "*ground truth value*" $E_{\mathcal{D}}[f]$. The user has no *direct access* to the sample $\bar{x}$. That is, he can *only* acquire information regarding $\bar{x}$ from the confidence intervals $\tilde{E}_{\bar{x}}[f] \pm \epsilon$ for the expectation $E_{\mathcal{D}}[f]$, which the testing procedure has returned as answer to the queries considered so far. Let $\mathcal{F}_k = \{f_1, \ldots, f_k\}$ denote the set of the first k functions evaluated during the adaptive process. The maximum error in estimating the expectations of the k functions is given by:

$$\Psi(\mathcal{F}_k, \bar{x}) = \sup_{f \in \mathcal{F}_k} | \frac{1}{m} \sum_{i=1}^m f(x_i) - E_{\mathcal{D}}[f] |$$
$$= \sup_{f \in \mathcal{F}_k} | \tilde{E}_{\bar{x}}[f] - E_{\mathcal{D}}[f] |.$$

In this work, we use the expression "*overfittig*" as follows: A given set of functions $\mathcal{F}_k$ is said to *overfit the sample* $\bar{x}$ if for any $f \in \mathcal{F}$ the value $\tilde{E}_{\bar{x}}[f]$ evaluated on $\bar{x}$ differs from the *true value* $E_{\mathcal{D}}[f]$ by more than the user given threshold ϵ. Our adaptive testing process halts at the first k-th step for which for which it cannot guarantee that the probability of overfitting is at most δ, that is, when $\Pr(\Psi(\mathcal{F}_k, \bar{x}) \leq \epsilon) \geq 1 - \delta$.

The process is fully adaptive. The choice of the function f_{k+1} evaluated at the $k + 1$-th step may depend on the information obtained during the first k steps. We make no assumptions on the processes according to which the functions are adaptively chosen to be tested, nor do we require the total number of tests to be fixed in advance. For simplicity, we assume that all functions are in the range $[0, 1]$. More general settings are discussed later in the paper.

Bounding the generalization error for the iterative process: The sequence of answers to the queries, $\tilde{E}_{\bar{x}}[f_1] \pm \epsilon, \tilde{E}_{\bar{x}}[f_2] \pm \epsilon, \ldots, \tilde{E}_{\bar{x}}[f_k] \pm \epsilon$ defines a filtration $\mathcal{L} = \{\mathcal{D}_k\}_{k \geq 0}$, such that

$$\mathcal{D}_0 = \mathcal{D} \quad \text{and} \quad \mathcal{D}_k = \{\mathcal{D} \mid \tilde{E}_{\bar{x}}[f_1] \pm \epsilon \wedge \cdots \wedge \tilde{E}_{\bar{x}}[f_k] \pm \epsilon\}.$$

The k-th query is chosen with respect to, and is answered in the filtered distribution $\mathcal{D}_{k-1}$. Our first step in developing RADABOUND is to adapt the Rademacher Complexity results to an iterative, adaptive sequence of queries.

Let E_k denote the event that the answer to the k-th query was within ϵ of the correct value, that is, $E_k := |\tilde{E}_{\bar{x}}[f] - E_{\mathcal{D}}[f_k]| \leq \epsilon$. Then, $\Pr(\Psi(\mathcal{F}_k, \bar{x}) \leq \epsilon) = \Pr(\wedge_{i=1}^k E_k)$, and thus, in the filtration process,

$$\Pr_{\mathcal{L}}(\Psi(\mathcal{F}_k, \bar{x}) > \epsilon)$$
$$\leq \Pr_{\mathcal{D}_0}(\bar{E}_1) + \Pr_{\mathcal{D}_1}(\bar{E}_2) + \ldots + \Pr_{\mathcal{D}_{k-1}}(\bar{E}_k)$$
$$= \sum_{i=1}^k \frac{\Pr(\bar{E}_i \wedge (\wedge_{j=1}^{i-1} E_j))}{\Pr(\wedge_{j=1}^{i-1} E_j)}$$
$$\leq \frac{1 - \Pr(\wedge_{j=1}^k E_j)}{\Pr(\wedge_{j=1}^{k-1} E_j)}.$$

By the definition of the events E_j we thus have:

$$\Pr_{\mathcal{L}}(\Psi(\mathcal{F}_k, \bar{x}) > \epsilon) \leq \frac{1 - \Pr(\Psi(\mathcal{F}_k, \bar{x}) \leq \epsilon)}{\Pr(\Psi(\mathcal{F}_{k-1}, \bar{x}) \leq \epsilon)} \quad (1)$$

where $\Pr()$ with no subscript refers to probability in the un-filtered distribution $\mathcal{D}$.

The fact that the distribution of the generalization error in the adaptive case, $\Pr_{\mathcal{L}}(\Psi(\mathcal{F}_k, \bar{x}) > \epsilon)$, is related to the probability of an error in the non-adaptive case, $\Pr(\Psi(\mathcal{F}_k, \bar{x}) > \epsilon)$, is not surprising. In order to fit the sample differently than the original distribution $\mathcal{D}$, the process needs to detect a pattern whose frequency is considerably different in the sample compared to the actual distribution $\mathcal{D}$. However, the first query that observes such a pattern is chosen when the process has not yet observed a significant difference between the sample and the distribution. This is due to the fact that the process halts as soon as such difference is detected. Thus, the probability of overfitting in k queries is related to the probability that the sample gives a bad estimate for the correct value of one of the k queries in the non-adaptive case.

The challenge is to compute a tight bound to the probability $\Pr(\Psi(\mathcal{F}_k, \bar{x}) \leq \epsilon)$. We achieve this through two novel bounds on estimating the Rademacher Complexity of $\mathcal{F}_k$.

III. BOUNDING $\Psi(\mathcal{F}_k, \bar{x})$ USING RADEMACHER COMPLEXITY

Our solution is based on iterative applications of Rademacher Complexity bounds.

Definition 1. *[19] Let* $\bar{\sigma} = (\sigma_1, \ldots, \sigma_m)$ *be a vector of* m *independent* Rademacher random variables, *such that for all* i, $Pr(\sigma_i = 1) = Pr(\sigma_i = -1) = 1/2$. *The* Empirical Rademacher Complexity *of a class of function* $\mathcal{F}$ *with respect to a sample* $\bar{x} = \{x_1, \ldots, x_m\}$, *with* $\bar{x} \sim \mathcal{D}^m$ *is*

$$R_{\bar{x}}^{\mathcal{F}} = E_{\bar{\sigma}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m f(x_i)\sigma_i \right]$$

The Rademacher Complexity *of* $\mathcal{F}$ *for samples of size* m *is defined as* $R_m^{\mathcal{F}} := E_{\bar{x} \sim \mathcal{D}^m}[R_{\bar{x}}^{\mathcal{F}}]$.

The relation between $\Psi(\mathcal{F}_k, \bar{x})$ and the Rademacher Complexity of $\mathcal{F}_k$ is given by the following results [1]:

Lemma 1 (*Lemma 26.2,* [4])**.**

$$E_{\bar{x} \sim \mathcal{D}^m}\left[\Psi(\mathcal{F}_k, \bar{x})\right] = E\left[\sup_{f \in \mathcal{F}_k} |\frac{1}{m}\sum_{i=1}^{m} f(x_i) - E_{\mathcal{D}}[f]|\right]$$
$$\leq 2R_m^{\mathcal{F}_k}.$$

Lemma 2 (*Theorem 14.21,* [19])**.** *Assume that for all $x \in \mathcal{X}$ and $f \in \mathcal{F}_k$ we have $f(x) \in [0,1]$, then:*

$$Pr\left(\Psi(\mathcal{F}_k, \bar{x}) > 2R_m^{\mathcal{F}_k} + \epsilon\right) \leq e^{-2m\epsilon^2}. \quad (2)$$

Note that in our context (a) we need a one-sided bound, and (b) for all $x \in \mathcal{X}$ and $f \in \mathcal{F}_k$ we have $f(x) \in [0,1]$.

For algorithmic applications, two important consequences of these result are that: (a) for bounded functions the generalization error is concentrated around their expectation, and (b) the Rademacher Complexity can be estimated from the sample. In order for this bound to be actually usable in practical applications, it is necessary to compute an estimate of the Rademacher Complexity given the dataset $\bar{x}$, and to bound its error. In the "*textbook*" treatment, the difference between Rademacher Complexity and its empirical counterpart is bounded using a second application of *McDiarmid's Inequality* [4], [20]. However, this bound is often too loose for practical applications such as ours.

In this work, we propose an alternative, *direct*, estimate of the Rademacher Complexity and we develop two methods for tightly bounding the estimation error.

A. Tight bounds on Rademacher Complexity estimate

Given a finite size sample $\bar{x} \sim \mathcal{D}^m$ and ℓ independent Rademacher vectors $\bar{\sigma}_1, \ldots, \bar{\sigma}_\ell$, each composed of m independent Rademacher random variables (i.e., $\bar{\sigma}_j = \sigma_{j,1}\sigma_{i,2}\ldots\sigma_{j,m}$), we estimate $R_m^{\mathcal{F}_k}$ with

$$\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} = \frac{1}{\ell}\sum_{j=1}^{\ell}\sup_{f \in \mathcal{F}_k}\frac{1}{m}\sum_{i=1}^{m}f(x_i)\sigma_{j,i}. \quad (3)$$

Clearly, $E_{\bar{x},\bar{\sigma}_1,\ldots,\bar{\sigma}_\ell}\left[\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}\right] = R_m^{\mathcal{F}_k}$. To bound the error $R_m^{\mathcal{F}_k} - \tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$, we model the process as a *Doob martingale* ([19, Chapter 13.1]) as follows:

$$C_i = E[R_m^{\mathcal{F}_k} - \tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} \mid Y_1, \ldots, Y_i] \ \text{ for } i = 0, \ldots, m(\ell+1),$$

where the $Y_1, \ldots, Y_{m(\ell+1)}$ are the random variables that determinate the value of the estimate $\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$. The first m variables Y_i's correspond to the values of the sample $\bar{x}$, i.e. for $1 \leq i \leq m$, $Y_i = X_i$, and the remaining $m\ell$ Y_i's correspond to the Rademacher random variables, $Y_i = \sigma_{\lfloor i/m \rfloor, i - \lfloor i/m \rfloor}$. It is easy to verify that $C_0 = 0$, and $C_{m(\ell+1)} = R_m^{\mathcal{F}} - \tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$.

Next, we define a *martingale difference sequence* $Z_i = C_i - C_{i-1}$ with respect to the martingale $C_0, C_1, \ldots C_{m(\ell+1)}$,

---

[1]In our setting, in order apply the result with absolute value we assume that for any $f \in \mathcal{F}_k$ we also have $-f \in \mathcal{F}_k$, i.e., we assume that $\mathcal{F}_k$ is *closed under negation*.

and note that $\sum_{t=1}^{m(\ell+1)} Z_t = C_{m(\ell+1)} = R_m^{\mathcal{F}_k} - \tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$.

Application of Bernstein's Inequality for Martingales: Our first bound builds on Bernstein's Inequality for Martingales (BIM). We use the following version due to Freedman [8], as presented in [21] and adapted to one-sided error.

Theorem 1. *[8], [21] Let $Z_1, \ldots, Z_t$ be a martingale difference sequence with respect to a certain filtration $\{\mathscr{F}_i\}_{i=0,\ldots,t}$. Thus, $E[Z_i | \mathscr{F}_{i-1}] = 0$ for $i = 1, \ldots, t$. The process $\sum_{i=1}^{t} Z_i$ is thus a martingale with respect to this filtration. Further, assume that $|Z_i| \leq a$ for $i = 1, \ldots, t$, and that the conditional variance $\sum_{i=1}^{t} E[Z_i^2] \leq L$. For $\epsilon \in (0,1)$, we have:*

$$Pr\left(\sum_{i=1}^{t} Z_i > \epsilon\right) \leq e^{-\frac{\epsilon^2}{2L + 2a\epsilon/3}}. \quad (4)$$

Note that the bound presented here is slightly different from the one in [8] as for our purposes we only require a one-sided bound. A careful analysis of $E[Z_i^2]$ in our application allows us to obtain a significantly stronger bound than the one obtained using McDiarmid's Inequality [4], [20], which depends on the maximum variation of the martingale.

Theorem 2. *Given a sample $\bar{x} \sim \mathcal{D}^m$, a family of functions $\mathcal{F}_k$ which take values in $[0,1]$, ℓ independent vectors of Rademacher random variables, and $\epsilon, \delta \in (0,1)$, we have:*

$$Pr\left(R_m^{\mathcal{F}_k} - \tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} > \epsilon\right) \leq e^{-\frac{6m\ell\epsilon^2}{15 + 8\ell\epsilon}}. \quad (5)$$

Proof: Recall the definition of the Doob martingale

$$C_i = E[R_m^{\mathcal{F}_k} - \tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} \mid Y_1, \ldots, Y_i],$$

for $i = 0, \ldots, m(\ell+1)$, and the the definition of the corresponding martingale difference sequence $Z_i = C_i - C_{i-1}$.

By definition, for every $i = 1, \ldots, m(\ell+1)$, we have $E[Z_i] = 0$, and hence, $E[Z_i^2] = \text{Var}[Z_i]$.

In order to apply Bernstein's Inequality, we need a bound a, such that $a \geq |Z_i|$ for $1 \leq i \leq m(\ell+1)$, and an upper-bound L to the conditional variance, such that

$$L \geq \sum_{i=1}^{m(\ell+1)} E[Z_i^2] = \sum_{i=1}^{m(\ell+1)} \text{Var}[Z_i].$$

We consider the cases for $1 \leq i \leq m$ and $m < i \leq m(\ell+1)$ separately:

- $1 \leq i \leq m$: For $1 \leq j \leq \ell$, let us consider

$$C_i^{(j)} = E[R_m^{\mathcal{F}_k} - \sup_{f \in \mathcal{F}_k}\frac{1}{m}\sum_{i=1}^{m}f(x_i)\sigma_{j,i} \mid Y_1, \ldots, Y_i],$$

$$Z_i^{(j)} = C_i^{(j)} - C_{i-1}^{(j)}.$$

According to our definitions, we have

$$C_i = \frac{1}{\ell}\sum_{j=1}^{\ell} C_i^{(j)}$$

$$Z_i = \frac{1}{\ell}\sum_{j=1}^{\ell} Z_i^{(j)}.$$

74

Since $\forall x \in \mathcal{X}$ and $\forall f \in \mathcal{F}_k$, $f(x) \in [0,1]$, changing the value of any of the m points in $\bar{x}$ can change $\sup_{f \in \mathcal{F}_k} \frac{1}{m} \sum_{i=1}^{m} f(x_i)$ by at most $1/m$, and thus we have $|Z_i^{(j)}| \leq 1/m$, and $Z_i^{(j)} \in [\alpha, \beta]$ with $\beta - \alpha \leq 1/m$. From *Popoviciu's Inequality on variance* [22], we have that the variance of a random variable which takes values in $[\alpha, \beta]$ is bounded from above by $(\beta - \alpha)^2/4$. Hence, by applying Popoviciu's Inequality to $Z_i^{(j)}$, we have that $\mathrm{Var}\left[Z_i^{(j)}\right] \leq 1/(4m^2)$.

As we are considering the expectation over the unassigned values of the Rademacher random variables, and as we are averaging over the values obtained using ℓ independent and identically distributed vectors of Rademacher random variables, we can conclude that $|Z_i| \leq \frac{1}{\ell} \sum_{j=1}^{\ell} |Z_i^{(j)}| \leq 1/m$, and $\mathrm{Var}[Z_i] = \frac{1}{\ell^2} \sum_{j=1}^{\ell} \mathrm{Var}\left[Z_i^j\right] \leq 1/(4m^2\ell)$.

- $m < i \leq m(\ell+1)$: Changing the value of any of the ℓm Rademacher random variables can change the value of $\tilde{R}_{\bar{x},j}^{\mathcal{F}_k}$ by at most $2/\ell m$, and thus we have $|Z_i| \leq 2/\ell m \leq 1/m$, and $Z_i \in [\alpha, \beta]$ with $\beta - \alpha \leq 2/\ell m$. By applying Popoviciu's Inequality, we thus have $\mathrm{Var}[Z_i] \leq 1/\ell^2 m^2$.

By linearity of expectation, $\sum_{i=1}^{m(\ell+1)} Z_i = R_m^{\mathcal{F}_k} - \tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$. Further, we have $\sum_{i=1}^{m(\ell+1)} Z_i^2 \leq 5/4\ell m$, and $|Z_i| < 1/m$ for all $1 \leq i \leq m(\ell+1)$. The statement follows by applying Theorem 1. ∎

Note that for a sufficiently large (constant) ℓ, the term $6\epsilon\ell$ dominates the denominator of the exponent in the right hand side of (5), giving a *fast rate of convergence* for the estimate. Our estimate fully characterizes the benefit achieved using multiple independent vectors of Rademacher random variables in estimating $\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$.

Combining the results of Theorem 2, Lemma 1, and Lemma 2 using the union bound, we obtain an empirical bound on $\Psi(\mathcal{F}_k, \bar{x})$.

Theorem 3. *Given a sample $\bar{x} \sim \mathcal{D}^m$, a family of functions $\mathcal{F}_k$ which take values in $[0,1]$, ℓ independent vectors of Rademacher random variables, and $\epsilon, \delta \in (0,1)$, we have:*

$$Pr\left(\Psi(\mathcal{F}_k, \bar{x}) > 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} + \epsilon\right) < \min_{\alpha \in (0,\epsilon)} e^{-2m(\epsilon-\alpha)^2} + e^{-\frac{3m\ell\alpha^2}{30+8\ell\alpha}}.$$
$$(6)$$

Proof: From Lemmas 1 and 2, we have:

$$\Pr\left(\Psi(\mathcal{F}_k, \bar{x}) > 2R_m^{\mathcal{F}_k} + \epsilon_1\right) \leq e^{-2m\epsilon_1^2}.$$

Theorem 2 characterizes the quality of the estimate of the Rademacher Complexity given by $\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$, computed as specified in (3):

$$\Pr\left(R_m^{\mathcal{F}} - \tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} > \epsilon_2\right) \leq e^{-\frac{6m\ell\epsilon_2^2}{15+8\ell\epsilon_2}}.$$

Combining the two results, we obtain:

$$\Pr\left(\Psi(\mathcal{F}_k, \bar{x}) > 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} + \epsilon_1 + 2\epsilon_2\right) \leq e^{-2m\epsilon_1^2} + e^{-\frac{6m\ell\epsilon_2^2}{15+8\ell\epsilon_2}}.$$

By substituting $\alpha = 2\epsilon_2$ and $\epsilon = \epsilon_1 + 2\epsilon_2$ in the previous equation, we have:

$$\Pr\left(\Psi(\mathcal{F}_k, \bar{x}) > 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} + \epsilon\right) \leq e^{-2m(\epsilon-\alpha)^2} + e^{-\frac{6m\ell(\alpha/2)^2}{15+8\ell\epsilon_2}}.$$

The statement follows. ∎

Alternative bound with single application of Bernstein's Inequality for Martingales: We now present an alternative result to the one in Theorem 3, which can be achieved with a single application of BIM. This bound is tighter than the one in Theorem 3 when the number of independent vectors of Rademacher random variables is very high.

Theorem 4.

$$Pr\left(\Psi(\mathcal{F}_k, \bar{x}) > 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} + \epsilon\right) < e^{-\frac{\epsilon^2}{\frac{\ell+4\sqrt{\ell}+20}{2m\ell} + \frac{4\epsilon}{3m}}}$$

Proof: Consider the Doob supermartingale:

$$C_i = \mathrm{E}\left[\Psi(\mathcal{F}_k, \bar{x}) - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} | Y_1, \dots Y_i\right] \quad \text{for } i = 0, \dots, m(\ell+1),$$

where for $1 \leq i \leq m$, $Y_i = X_i$, and the remaining Y_i correspond to the $m\ell$ independent Rademacher random variables in the ℓ vectors; that is, $Y_{j(m)+i} = \sigma_{j,i}$ for $1 \leq j \leq \ell$ and $1 \leq i \leq m$. It is easy to verify that $C_{m(\ell+1)} = \Psi(\mathcal{F}_k, \bar{x}) - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$. Further, $C_0 = \mathrm{E}\left[\Psi(\mathcal{F}_k, \bar{x})\right] - 2R_m^{\mathcal{F}_k}$, and due to Theorem 1, $C_0 \leq 0$.

Let us define the corresponding martingale difference sequence $Z_i = C_i - C_{i-1}$. For each $i \in \{1, \dots, m(\ell+1)\}$, due to linearity of expectation, we have $Z_i = A_i - 2B_i$, where:

$$A_i = \mathrm{E}\left[\Psi(\mathcal{F}_k, \bar{x}) | Y_1, \dots Y_i\right] - \mathrm{E}\left[\Psi(\mathcal{F}_k, \bar{x}) | Y_1, \dots Y_{i-1}\right];$$
$$B_i = \mathrm{E}\left[\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} | Y_1, \dots Y_i\right] - \mathrm{E}\left[\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} | Y_1, \dots Y_{i-1}\right].$$

In order apply Bernstein's Inequality, we need an upper-bound $a \geq |Z_i|$ for $1 \leq i \leq m(\ell+1)$ and an upper-bound L, such that $L \geq \sum_{i=1}^{m(\ell+1)} \mathrm{E}\left[Z_i^2\right]$.

Given our definition of Z_i, we have that for every i, $\mathrm{E}[Z_i] = \mathrm{E}[A_i] = \mathrm{E}[B_i] = 0$, and thus: $\mathrm{E}\left[Z_i^2\right] = \mathrm{Var}[Z_i] \leq \mathrm{Var}[A_i] + 4\mathrm{Var}[B_i] + 4\mathrm{Cov}[A_i, B_i]$. From the properties of covariance, we have $|\mathrm{Cov}[A_i, B_i]| \leq \sqrt{\mathrm{Var}[A_i]\mathrm{Var}[B_i]}$, and thus, $\mathrm{E}\left[Z_i^2\right] = \mathrm{Var}[Z_i] \leq \mathrm{Var}[A_i] + 4\mathrm{Var}[B_i] + 4\sqrt{\mathrm{Var}[A_i]\mathrm{Var}[B_i]}$.

We consider the cases for $1 \leq i \leq m$ and $m < i \leq m(\ell+1)$ separately:

- $1 \leq i \leq m$: In our setting $\forall x \in \mathcal{X}$ and $\forall f \in \mathcal{F}$, $f(x) \in [0,1]$, changing the value of any of the m points in $\bar{x}$ can change $f(\bar{x})$ by at most $1/m$. Therefore, $|A_i| \leq 1/m$, and $A_i \in [\alpha, \beta]$ with $\beta - \alpha \leq 1/m$. By applying Popoviciu's Inequality, we have: $\mathrm{Var}[A_i] \leq 1/4m^2$.

The analysis for $\mathrm{Var}[B_i]$ follows the same reasoning discussed in the proof of Theorem 2 for bounding $\mathrm{Var}[Z_i]$

in the case $1 \leq i \leq m$, and thus $\text{Var}\left[B_i\right] \leq 1/(4m^2\ell)$. We can thus conclude:

$$\text{E}\left[Z_i^2\right] = \text{Var}\left[Z_i\right] \leq \frac{1}{4m^2} + \frac{4}{4\ell m^2} + 4\sqrt{\frac{1}{4m^2}\frac{1}{4\ell m^2}}$$
$$\leq \frac{\ell + 4 + 4\sqrt{\ell}}{4m^2\ell};$$
$$|Z_i| \leq \frac{2}{m}.$$

- $m < i \leq m(\ell+1)$: Changing the value of any of the Rademacher random variables does not change the value of $\Psi\left(\mathcal{F}_k, \bar{x}\right)$. Hence, $\text{Var}\left[A_i\right] = \text{E}\left[A_i^2\right] = 0$.

  Given fixed values for the random variables corresponding to the points in $\bar{x}$, changing the value of one Rademacher random variable can change the value of $\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$ by at most $2/\ell m$. Thus, $|B_i| \leq \frac{2}{\ell m}$, and $B_i \in [\alpha, \beta]$ with $\beta - \alpha \leq 2/\ell m$. By applying Popoviciu's Inequality, we have:

$$\text{Var}\left[Z_i\right] = 4\text{Var}\left[B_i\right] \leq \frac{4}{m^2\ell^2}.$$

We, therefore, have $|Z_i| \leq \frac{2}{m}$ for all $1 \leq i \leq m(\ell+1)$, and $\sum_{i=1}^{m(\ell+1)} \text{E}\left[Z_i^2\right] \leq \frac{\ell+4\sqrt{\ell}+20}{4m\ell}$. By linearity of expectation, and by applying Theorem 1:

$$\sum_{i=1}^{\ell(m+1)} Z_i = \Psi\left(\mathcal{F}_k, \bar{x}\right) - \text{E}\left[\Psi\left(\mathcal{F}_k, \bar{x}\right)\right] - 2\left(\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} - R_m^{\mathcal{F}_k}\right)$$
$$\geq \Psi\left(\mathcal{F}_k, \bar{x}\right) - 2R_m^{\mathcal{F}_k} - 2\left(\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} - R_m^{\mathcal{F}_k}\right);$$
$$\geq \Psi\left(\mathcal{F}_k, \bar{x}\right) - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k};$$

The statement follows by applying BIM (Theorem 3.3). ∎

This result can be used in RADABOUND in place of the bound given by Theorem 3. Note that with this result, it is easier to compute the bound on the probability of overfitting (denoted as δ' in line 11: of Algorithm 1).

Application of the Martingale Central Limit Theorem: In practical applications, one may prefer the standard practice in statistics of applying central limit asymptotic bounds. We develop here a bound based on the Martingale Central Limit Theorem (MCLT). Our experimental results in Section V show that the bound obtained using the MCLT is more powerful while still preserving statistical validity.

We adapt the following version of the MCLT [2]:

Theorem 5 (*Corollary 3.2, [9]*). *Let $Z_0, Z_1, \ldots$ be a difference martingale with bounded absolute increments. Assume that (1) $\sum_{i=1}^n Z_i^2 \overset{p}{\to} V^2$ for a finite $V > 0$, and (2) $\text{E}\left[\max_i Z_i^2\right] \leq M < \infty$, then $\sum_{i=1}^n Z_i / \sqrt{\sum_{i=1}^n E\left[Z_i^2\right]}$ converges in distribution to $N(0,1)$.*

When applying the MCLT, there is no advantage in bounding separately $\Psi\left(\mathcal{F}_k, \bar{x}\right) - 2R_m^{\mathcal{F}_k}$ and $2R_m^{\mathcal{F}_k} - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$. Instead,

[2]Formally, the asymptotic is defined on a triangle array, where rows are samples of growing sizes. We also assume that all expectations are well-defined in the corresponding filtration.

we compute a bound on the distribution of $\Psi\left(\mathcal{F}_k, \bar{x}\right) - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$ by analyzing the *Doob supermartingale*

$$C_i = E[\Psi\left(\mathcal{F}_k, \bar{x}\right) - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} \mid Y_1, \ldots, Y_i]$$

for $i = 0, \ldots, m(\ell+1)$, with respect to the same $Y_1, \ldots, Y_{m(\ell+1)}$ defined as in Section III-A.

As in the finite sample case, the following theorem relies on a careful analysis of $\text{E}\left[Z_i^2\right]$ for the martingale difference sequence $Z_i = C_i - C_{i-1}$.

Theorem 6. *Given a sample $\bar{x} \sim \mathcal{D}^m$, a family of functions $\mathcal{F}_k$ which take values in $[0,1]$, ℓ independent vectors of Rademacher random variables, and $\epsilon, \delta \in (0,1)$, we have:*

$$\lim_{m \to \infty} Pr\left(\Psi\left(\mathcal{F}_k, \bar{x}\right) - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} > \epsilon\frac{\sqrt{\ell + 4\sqrt{\ell} + 20}}{2\sqrt{\ell m}}\right) < 1 - \Phi\left(\epsilon\right).$$

Where $\Phi(x)$ denotes the cumulative distribution function for the standard normal distribution.

Proof: The proof closely follows the steps of the proof of Theorem 4. Consider the Doob supermartingale for the function $\Psi\left(\mathcal{F}_k, \bar{x}\right) - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}$:

$$C_i = \text{E}\left[\Psi\left(\mathcal{F}_k, \bar{x}\right) - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k}|Y_1, \ldots Y_i\right] \text{ for } i = 0, \ldots, m(\ell+1),$$

where for $1 \leq i \leq m$, $Y_i = X_i$, and the remaining Y_i correspond to the $m\ell$ independent Rademacher random variables in the ℓ vectors. That is, $Y_{j(m)+i} = \sigma_{j,i}$, for $1 \leq j \leq \ell$ and $1 \leq i \leq m$. Further, let us define the corresponding martingale difference sequence $Z_i = C_i - C_{i-1}$.

In order to apply the MCLT, we need to bound $\sum_{i=1}^{m(\ell+1)} \text{E}\left[Z_i^2\right]$ from above, and we need to verify that $|Z_i|$ is bounded.

Note that the sequence Z_i defined here corresponds to the martingale difference sequence by the same name that we studied in the proof of Theorem 4. As shown in the proof of Theorem 4, we have $\sum_{i=1}^{m(\ell+1)} \text{E}\left[Z_i^2\right] \leq \frac{\ell+4\sqrt{\ell}+20}{4m\ell}$, and $|Z_i| \leq 2/m$ for all $1 \leq i \leq m(\ell+1)$.

Applying the MCLT, we have that as m goes to infinity, $\sum_{i=1}^{m(\ell+1)} Z_i / \sqrt{\sum_{i=1}^{m(\ell+1)} \text{E}\left[Z_i^2\right]}$ converges in distribution to $N(0,1)$, and thus:

$$\lim_{m \to \infty} \text{Pr}\left(\sum_{i=1}^{\ell(m+1)} Z_i \left(\sqrt{\sum_{i=1}^{m(\ell+1)} \text{E}\left[Z_i^2\right]}\right)^{-1} > \epsilon\right) < 1 - \Phi\left(\epsilon\right),$$

$$\lim_{m \to \infty} \text{Pr}\left(\sum_{i=1}^{\ell(m+1)} Z_i > \epsilon\sqrt{\frac{\ell + 4\sqrt{\ell} + 20}{4m\ell}}\right) < 1 - \Phi\left(\epsilon\right),$$

By linearity of expectation, and by applying Theorem 1:

$$\sum_{i=1}^{\ell(m+1)} Z_i = \Psi\left(\mathcal{F}_k, \bar{x}\right) - \text{E}\left[\Psi\left(\mathcal{F}_k, \bar{x}\right)\right] - 2\left(\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} - R_m^{\mathcal{F}_k}\right)$$
$$\geq \Psi\left(\mathcal{F}_k, \bar{x}\right) - 2R_m^{\mathcal{F}_k} - 2\left(\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k} - R_m^{\mathcal{F}_k}\right);$$
$$\geq \Psi\left(\mathcal{F}_k, \bar{x}\right) - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_k};$$

The statement follows. ∎

76

Due to its asymptotic nature, it is not possible to compare directly the tightness of the bound in Theorem 6 with that of finite sample bounds such as the one in Theorem 3. Still, this bound is of great interest in many practical scenarios as it allows for a much tighter bound for the generalization error.

IV. THE RADABOUND ALGORITHM

The algorithm starts by drawing ℓ *independent* vectors of Rademacher variables. These vectors are fixed throughout the execution of the algorithm. The advantage of fixing the Rademacher vectors is that (1) we deal with a nested sequence of events, $\mathcal{F}_{k-1} \subseteq \mathcal{F}_k$, and (2) the actual computation of the Rademacher complexity estimate is simple and efficient.

Computing the estimate: At the end of each round k, the algorithm stores for each of the Rademacher vectors $j = 1, \ldots, \ell$, the value $\tilde{M}_{\bar{x},j}^{\mathcal{F}_k} = \max_{f \in \mathcal{F}_k} \frac{1}{|\bar{x}|} \sum_{i=1}^{|\bar{x}|} f(x_i)\sigma_{i,j}$. To update these values, at iteration $k+1$, the algorithm computes

$$\tilde{M}_{\bar{x},j}^{\mathcal{F}_{k+1}} \leftarrow \max\{\tilde{M}_{\bar{x},j}^{\mathcal{F}_k}, \frac{1}{|\bar{x}|} \sum_{i=1}^{m} f_{k+1}(x_i)\sigma_{i,j}\}, \quad j = 1, \ldots, \ell.$$

The estimate of the Rademacher Complexity at round $k+1$ is then given by $\tilde{R}_{\bar{x}}^{\mathcal{F}_{k+1}} = \frac{1}{\ell} \sum_{j=1}^{\ell} \tilde{M}_{\bar{x},j}^{\mathcal{F}_{k+1}}$.

ALGORITHM 1 RADABOUND - Adaptive data analysis with Rademacher Complexity control

---

1: **procedure** RADABOUND($\bar{x}, \varepsilon, \delta, \ell$)
2:   $m \leftarrow |\bar{x}|$ ▷ Size of the input sample
    ▷ Initialization estimator for Rademacher Complexity
3:   **for** $j \in \{0, 1, \ldots, \ell\}$ **do**
4:      $\sigma_j \leftarrow$ vector of m iid Rademacher RVs
5:      $R_{\bar{x},j}^{\mathcal{F}_0} \leftarrow 0$
    ▷ Main execution body
6:   **while** new k-th query f_k from the stream **do**
7:      $\mathcal{F}_{k+1} \leftarrow \mathcal{F}_k \cup \{f_k\}$
       ▷ Rademacher Average estimation update
8:      **for** $j \in \{0, 1, \ldots, \ell\}$ **do**
9:         $R_{\bar{x},j}^{\mathcal{F}_{k+1}} \leftarrow \max\{R_{\bar{x},j}^{\mathcal{F}_k}, \frac{1}{m} \sum_{i=1}^{m} f_k(x_i)\sigma_{j,i}\}$
10:       $\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_{k+1}} \leftarrow \frac{1}{\ell} \sum_{j=1}^{\ell} R_{\bar{x},j}^{\mathcal{F}_{k+1}}$
       ▷ Control with BIM
11:      $\delta' \leftarrow e^{-\left(\min 0, \epsilon - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_{k+1}}\right)^2 / \frac{\ell + 4\sqrt{\ell} + 20}{2m\ell} + \frac{4\epsilon}{3m}}$
       ▷ Control with MCLT- Alternative to 11:
12:      **or** $\delta' \leftarrow 1 - \Phi\left(\max\{0, \epsilon - 2\tilde{R}_{\bar{x},\ell}^{\mathcal{F}_{k+1}}\} \sqrt{\frac{4\ell m}{\ell + 4\sqrt{\ell} + 20}}\right)$
       ▷ Overfit control test
13:      **if** $\delta' \leq \delta(1-\delta)$ **then**
14:         **return** $\frac{1}{m} \sum_{x \in \bar{x}} f(x)$
15:      **else**
16:         **Halt: Cannot guarantee the statistical validity of further queries.**

Stopping rule: Given real values $\epsilon, \delta \in (0, 1)$, the procedure halts at the first k-th step for which it cannot guarantee that

$$\Pr_{\mathcal{L}}\left(\Psi(\mathcal{F}_{k+1}, \bar{x}) > \epsilon\right) \leq \delta.$$

Recall from (1) that

$$\Pr_{\mathcal{L}}\left(\Psi(\mathcal{F}_k, \bar{x}) > \epsilon\right) \leq \frac{\Pr\left(\Psi(\mathcal{F}_k, \bar{x}) > \epsilon\right)}{\Pr\left(\Psi(\mathcal{F}_{k-1}, \bar{x}) \leq \epsilon\right)}. \quad (7)$$

Since $\Pr(\Psi(\mathcal{F}_k, \bar{x}) > \epsilon) \geq \Pr(\Psi(\mathcal{F}_{k-1}, \bar{x}) > \epsilon)$, it is sufficient to require $\Pr(\Psi(\mathcal{F}_k, \bar{x}) > \epsilon) < \delta(1-\delta)$ to have $\Pr_{\mathcal{L}}\left(\Psi(\mathcal{F}_k, \bar{x}) > \epsilon\right) \leq \delta$, and we can use the bounds obtained in Theorem 3 or Theorem 6. Thus, we prove

Theorem 7. *Given a sample $\bar{x} \sim \mathcal{D}^m$, let $\mathcal{F}_k$ denote the set of functions adaptively selected during the first k steps. If* RADABOUND *has not halted at step k, then*

$$Pr_{\mathcal{L}}\left(\Psi(\mathcal{F}_k, \bar{x}) \leq \epsilon\right) > 1 - \delta.$$

The bound in Theorem 6 based on the MCLT can be used in RADABOUND *as an alternative* to the bound in Theorem 3 (lines 11-12 in Algorithm 1). In Section 5, we present an experimental comparison of performance of RADABOUND when using the two methods.
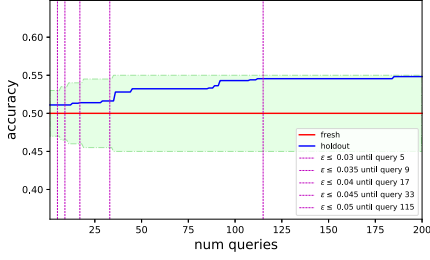
V. EXPERIMENTAL RESULTS

We demonstrate the power and efficiency of our technique through a variety of experiments. Our experimental setup is similar to the one used in the state-of-the-art [1], except that all our reported results are for ranges of parameters for which we actually have provable statistical guarantees.
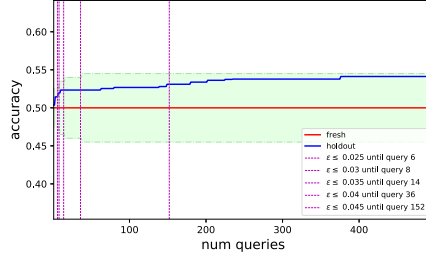
Experimental Setup: We consider a learning task of classifying vectors composed by d features to the classes "-1" or "1". We consider only linear classifier vectors $\mathbf{w} \in \{-1, 0, 1\}^d$, assigning vector $\mathbf{x}$ to class $\mathbf{h}(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x})$. The goal of the learning algorithm is to find a classifier with minimum expected loss for the $0, 1$ *hard loss function* (0 for correct classification, 1 otherwise). To model a typical learning scenario, the learning algorithm is given two independent datasets. A *training set* X_T and an *holdout set* X_H. We then evaluate the performance of the learning algorithm using a third, independent *fresh set*. Evaluation on the fresh, independent set provides a baseline for the actual performance of the classifier being considered on an independent dataset. The goal of the algorithm being tested if to obtain a classifier $h()$ which "*fits the data*" as best as possible, that is the a classifier which minimizes the *Empirical Risk* over the holdout data, that is $\frac{1}{m} \sum_{\mathbf{x} \in X_H} \ell(h(\mathbf{x}), y(\mathbf{x}))$, where $y(\mathbf{x})$ denotes the "*true label*" associated with the data point $\mathbf{x}$ and $\ell(h(\mathbf{x}), y(\mathbf{x}))$ denotes the loss function. In our setting we consider the $\{0, 1\}$ loss function such that $\ell(h(\mathbf{x}), y(\mathbf{x})) = 0$ if $h(\mathbf{x}), y(\mathbf{x})$ (i.e., the classifier $h()$ assigns the correct label to $\mathbf{x}$), or $\ell(h(\mathbf{x}), y(\mathbf{x})) = 1$ otherwise.

Our learning algorithm works as follows: In the first phase, the algorithm evaluates the correlation between the values of the features and the labels of the vectors using *only* the training dataset X_T as $c_i = \frac{1}{m} \sum_{\mathbf{x} \in X_T} \mathbf{x}[i] l(\mathbf{x})$. The features are then sorted (in descending order) according to the absolute values of their correlations $|c_i|$ to the labels.
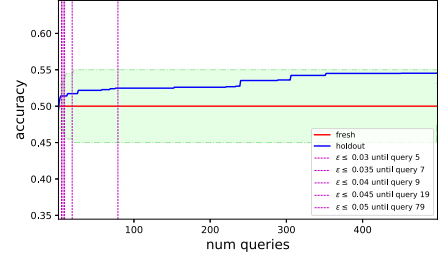
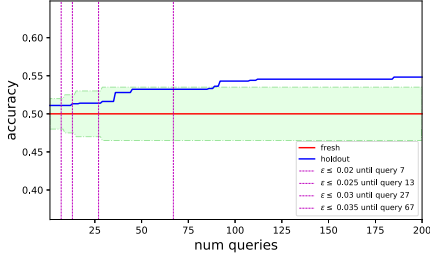The actual adaptive analysis of the data occurs in the second phase of the algorithm using the holdout data X_H.
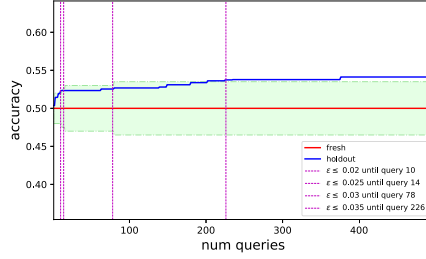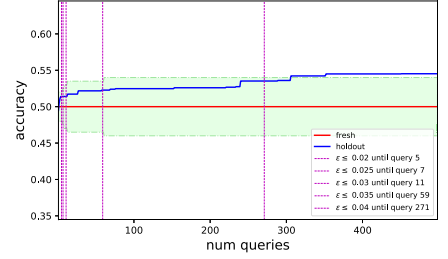
(a) Bernstein bound | (a) Bernstein bound | (a) Bernstein bound

(b) MCLT | (b) MCLT bound | (b) MCLT bound

Fig. 1: No signal. Feature values from $\mathcal{N}(0,1)$. $\delta = 0.1$.

Fig. 2: No signal. Feature values from $\mathcal{N}(0,2)$. $\delta = 0.15$.

Fig. 3: No signal. Feature values from $\mathcal{N}(0,8)$. $\delta = 0.2$.

The algorithm starts with a classifier $\mathbf{w} = 0$. It then considers features according to the order computed in the first phase. That is, features with *stronger* correlation (either positive or negative) to the label are considered first. Using the holdout set X_H, the algorithm tests, for each feature, whether assigning weight -1 or 1 to it improves the performance of the current best classifier. If that is the case, the classifier is updated with the new value for the feature; otherwise, the feature is left with weight zero. Each newly tested classifier is added to the class function $\mathcal{F}_k$. RADABOUND then computes a new estimate $\tilde{R}_m^{\mathcal{F}_k}$ of the Rademacher Complexity of $\mathcal{F}_k$, and uses it to determinate whether the total accumulated error is below ϵ with probability at least $1 - \delta$ as discussed in Section IV.
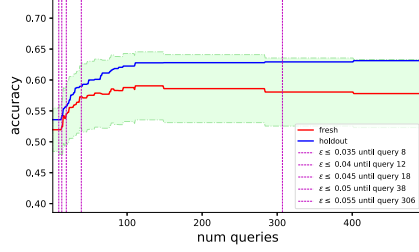
Data generation: In all the experiments $|X_T| = |X_H| = 4000$. Each vector in the dataset has 500 features. The estimation of the Rademacher Complexity $\tilde{R}_{\bar{x},\ell}^{\mathcal{F}}$ is computed according to (3) using $\ell = 32$ vectors of Rademacher random variables. We report results using (a) Bernstein's Inequality (Section III-A) and, (b) the MCLT (Section III-A). We consider the two following scenarios:

- **No signal in the data:** In this setting, each point $\mathbf{x} \in X_H$ is assigned a label independently and uniformly at random. The feature values are taken *independently* from a normal distribution with expectation 0 and various variance values. Thus, there is no correlation between the labels and the values of the features. We report the results in Figures 1-3.
- **Signal in the data:** In this setting the "*strength*" of the correlation between some features and the labels is characterized by two parameters: n, the number of the queries whose value is correlated to the label, and
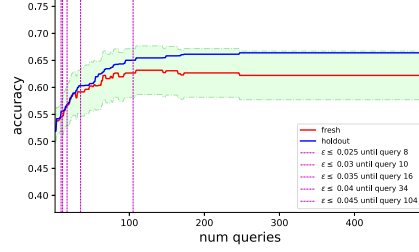
(positive or negative) $bias$ which defines the strength and sign of the correlation. We first generate datasets with no signal, like in the previous setting. We then fix a set of 50 features to be correlated with the label of their vectors. Letting $l(\mathbf{x})$ be the label of vector $\mathbf{x}$, the 50 correlated features of $\mathbf{x}$ are modified by adding $bias \times l(\mathbf{x})$ to their original value. We report the results in Figures 4-6.

Our experimental evaluation uses *synthetic data* generated randomly according to the previously discussed specifications. Besides being widespread in statistics literature and used in our most direct term of comparison [2], using synthetic data is particularly useful in this setting for two main reasons: (i) it allows evaluate the performance of our testing procedure while evaluating different settings by modifying its parameters (e.g., the size of the input sample, the strength of the signal being observed), and (ii), most importantly, it allows to verify whether/when overfitting (due to adaptivity) has *actually occurred* (i.e., when the blue line in the figures exceeds the green shaded region). This allows to asses both the *correctness* of the method and its *power*.
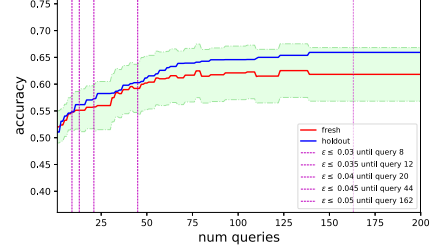
Results: The results of the experiments in the no signal (resp., signal in the data) setting are displayed in Figures 1-3 (resp., Figures 4-6). Each figure above corresponds to several runs of an experiment with the same parameters but different values of ϵ (the error bound). The blue line gives the accuracy of the best classifier computed after running the corresponding number of queries on the holdout set X_H. The red line gives the accuracy of the same classifier on fresh data (i.e., the red line represents the ground truth performance of the classifier being evaluated on a fresh sample). The vertical bars give the computed stopping time for each value of ϵ according to
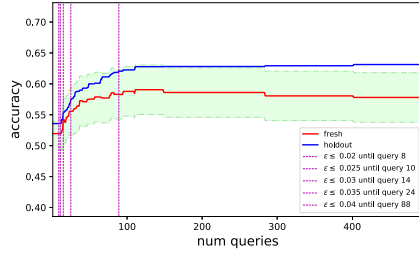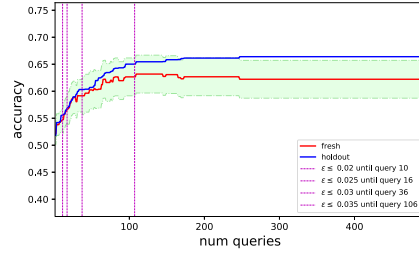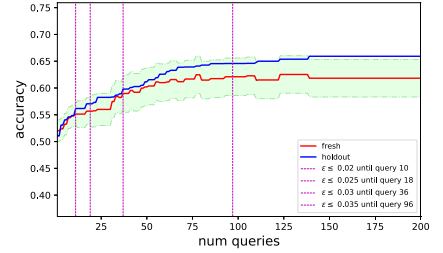
78

(a) Bernstein bound

(a) Bernstein bound

(a) Bernstein bound

(b) MCLT bound

(b) MCLT bound

(b) MCLT bound

Fig. 4: Signal. Feature values from $\mathcal{N}(0,4)$, $\delta = 0.1$, $bias = 0.5$.

Fig. 5: Signal. Feature values from $\mathcal{N}(0,2)$, $\delta = 0.15$, $bias = 0.5$.

Fig. 6: Signal. Feature values from $\mathcal{N}(0,2)$, $\delta = 0.2$, $bias = 0.25$.

the "*stopping rule*" of RADABOUND. The shaded green area corresponds to $\pm\epsilon$ values around the "*true accuracy*" of the classifier (the red line), for the ϵ value associated with the next vertical bar. The green shaded area beyond the last bar uses the same ϵ as the last bar. In a correct execution of RADABOUND, the blue line does not exit the green shaded area before the last vertical bar. The power of RADABOUND is measured by how close is the last bar to the first time the green line exit the shaded area.

The results of the experimental evaluations, as presented in the figures, demonstrate that RADABOUND successfully halts the sequence of tests before overfitting for the various values of ϵ, as the green line corresponding to the values of the function evaluated on the holdout does not exit the green shaded area before the corresponding vertical bar. The statistical power of the procedure is highlighted in particular by the result of experiments for which there is an actual correlation between the labels (Figures 4-6) and the value of the features as the overfitting control ensured by RADABOUND is not achieved at the expense of detecting the signal in the data.

In several scenarios (e.g., Figures 4 and 6), RADABOUND halts its execution very close to the first iteration for which overfit (with respect to the value of ϵ) actually occurs. RAD-ABOUND does not appear to be influenced by the distribution $\mathcal{D}$ over the data, but rather it behaves differently depending on the actual family of functions being tested.

For similar ϵ, δ parameters, the state-of-the-art `Thresholdout` algorithm [1] would require an holdout dataset of size $\sim 4 \times 10^6$ to provide answers to just 10 queries (details in Section VI). In contrast, our experiments show that RADABOUND can provably handle such parameters with a holdout set of just 4000 samples, thus with an improvement of almost three orders of magnitude in terms of *sample complexity*. The comparison is further discussed in Section VI. Using the MCLT leads to a tight analysis of $\Psi(\mathcal{F}_k, \bar{x})$, which in turn allows testing a higher number of adaptively chosen classifiers before halting and without overfitting (Figures 1-6b), compared to the stopping points obtained using the BIM (Figures 1-6a).

Finally, the fact that even when using the bounds obtained using the MCLT, the procedure halts correctly, further suggests that, despite their "asymptotic" nature, these are actually highly reliable even when dealing with an input sample of relatively small dimension.

VI. COMPARISON WITH METHODS BASED ON DIFFERENTIAL PRIVACY

The `Thresholdout` algorithm [1] provides guarantees similar to those of RADABOUND. `Thresholdout` operates using two datasets: a public dataset and a private *holdout* dataset. Every time a new query is received the algorithm evaluates its value on both the public and the private dataset. If their absolute difference is within a given *threshold*, `Thresholdout` returns to the user the value observed on the public dataset after perturbing it with some noise. Viceversa, if the absolute difference is higher than a certain threshold, the algorithm detects that the query being considered is overfitting on the public dataset. In this case, `Thresholdout` may instead provide the value computed on the private dataset after perturbing it with noise. As this last operation effectively "*leaks*" information regarding the holdout it can be executed up to B times, where B must be fixed prior to the execution of the algorithm.

To characterize the number of queries for which the `Thresholdout` algorithm provides provable statistical guarantees we apply Theorem 25 in [1] [3]. The theorem states that when testing up to k queries, with up to $B = 1$ of those being answered using the private "Holdout set", the size of the holdout set must be at least:

$$n \geq 96\epsilon^{-2} \ln\left(4k\delta^{-1}\right) \min\{80\sqrt{B\ln(1/\epsilon\delta)}, 16B\}$$

Thus, for $k = 10$, $B = 1$, $\epsilon = 0.5$, and $\delta = 0.1$: the required sample size is at least

$$n \geq 400 \times 96 \times \ln(400) \min\{80\sqrt{\ln(200)}, 16\} \geq 3.7 \times 10^6.$$

Therefore, even when requesting such, fairly loose, guarantees, `Thresholdout` requires an extremely high sample size in order to provide reliable answer to a handful of queries.

In contrast, we showed in Section V that RADABOUND can provably handle problems with these, and better, parameters while using a holdout set composed by just 4000 samples. Thus RADABOUND achieves an improvement of almost three orders of magnitude in terms of *sample complexity* compared to `Thresholdout`.

Further, `Thresholdout` requires that the user specifies *before* the execution the number of queries k which are going to be adaptively chosen to be tested, and the number B of maximum times that the algorithm can tolerate overfit on the public dataset by revealing information from the private "*holdout*" dataset. These requirements limit the adaptiveness of the process. In contrast RADABOUND uses the holdout dataset *as much as possible* without a fixed maximum number of queries. Finally, using Rademacher Complexity in evaluating the stopping criterion of the adaptive testing procedure allows RADABOUND to evaluate the properties of the actual family of functions tested so far (i.e., their *expressiveness*), rather than just its cardinality, in order to provide guarantees on the quality of the evaluations obtained in the adaptive analysis.

VII. CONCLUSION

We presented a rigorous, efficient and practical method for bounding the generalization error in an adaptive sequence of queries tested on the same dataset. While the standard "rule of thumb" for responsible data analysis and machine learning is to use a test set only once, our results demonstrate that, with an appropriate control mechanism, it may be possible to use the same test set more than once without significantly reducing the validity of the results. For concreteness, we focused here on the problem of evaluating the expectations of a set of functions in the range $[0, 1]$. This problem corresponds to the basic machine learning task of evaluating the correctness of classifiers with bounded $[0, 1]$ loss functions. We note that

our methods can be extended to a more general setting, for example using new concentration bounds on sub-exponential distributions [23], and self-bounding functions [24].

REFERENCES

[1] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *Advances in Neural Information Processing Systems*, 2015, pp. 2350–2358.

[2] ——, "The reusable holdout: Preserving validity in adaptive data analysis," *Science*, vol. 349, no. 6248, pp. 636–638, 2015.

[3] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth, "Preserving statistical validity in adaptive data analysis," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, 2015, pp. 117–126.

[4] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[5] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[6] P. L. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation," *Mach. Learn.*, vol. 48, pp. 85–113, 2002.

[7] V. Koltchinskii, "Rademacher penalties and structural risk minimization," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1902–1914, Jul. 2001.

[8] D. A. Freedman, "On tail probabilities for martingales," *the Annals of Probability*, pp. 100–118, 1975.

[9] P. Hall and C. C. Heyde, *Martingale limit theory and its application*. Academic press, 2014.

[10] C. E. Bonferroni, *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber, 1936.

[11] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.

[12] Y. Hochberg, "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.

[13] M. G. G'Sell, S. Wager, A. Chouldechova, and R. Tibshirani, "Sequential selection procedures and false discovery rate control," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 78, no. 2, pp. 423–444, 2016.

[14] D. P. Foster and R. A. Stine, "α-investing: a procedure for sequential control of expected false discoveries," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 2, pp. 429–444, 2008.

[15] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC*, vol. 3876. Springer, 2006, pp. 265–284.

[16] M. Hardt and J. Ullman, "Preventing false discovery in interactive data analysis is hard," in *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE, 2014, pp. 454–463.

[17] T. Steinke and J. Ullman, "Interactive fingerprinting codes and the hardness of preventing false discovery," in *Conference on Learning Theory*, 2015, pp. 1588–1628.

[18] A. Blum and M. Hardt, "The ladder: A reliable leaderboard for machine learning competitions," in *International Conference on Machine Learning*, 2015, pp. 1006–1014.

[19] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge university press, 2017.

[20] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Mar. 2003.

[21] K. Dzhaparidze and J. Van Zanten, "On bernstein-type inequalities for martingales," *Stochastic processes and their applications*, vol. 93, no. 1, pp. 109–117, 2001.

[22] T. Popoviciu, "Sur les èquations algèbriques ayant toutes leurs racines rèelles," *Mathematica*, vol. 9, pp. 129–145, 1935.

[23] A. Kontorovich, "Concentration in unbounded metric spaces and algorithmic stability," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 28–36.

[24] L. Oneto, A. Ghio, D. Anguita, and S. Ridella, "An improved analysis of the Rademacher data-dependent bound using its self bounding property," *Neural Networks*, vol. 44, pp. 107–111, 2013.

---

[3] A careful reader will notice that Figures 1-3 in [1] (the same figure appears as Figures 1 and 2 in [2]) represent an idealized illustration rather than statistically valid results. For the sample size used in these figures, the bound on the error probability of the threshold algorithm is not smaller than 1. Furthermore, the results crucially depend on a preprocessing of the two data sets (lines 95-97 in the "`runClassifier`" procedure in the python code in the supporting materials) that is not discussed in the paper.