
The Everlasting Database: Statistical Validity at a Fair Price

Blake Woodworth
Toyota Technological
Institute at Chicago

Vitaly Feldman
Google

Saharon Rosset
Tel Aviv University

Nathan Srebro
Toyota Technological
Institute at Chicago

Abstract

The problem of handling adaptivity in data analysis, intentional or not, permeates a variety of fields, including test-set overfitting in ML challenges and the accumulation of invalid scientific discoveries. We propose a mechanism for answering an arbitrarily long sequence of potentially adaptive statistical queries, by charging a price for each query and using the proceeds to collect additional samples. Crucially, we guarantee statistical validity without any assumptions on how the queries are generated. We also ensure with high probability that the cost for M non-adaptive queries is $O(\log M)$, while the cost to a potentially adaptive user who makes M queries that do not depend on any others is $O(pM)$.

1 Introduction

Consider the problem of running a server that provides the test loss of a model on held out data, e.g. for evaluation in a machine learning challenge. We would like to ensure that all test losses returned by the server are accurate estimates of the true generalization error of the predictors.

Returning the empirical error on held out test data would initially be a good estimate of the generalization error. However, an analyst can use the empirical errors to adjust their model and improve their performance on the test data. In fact, with a number of queries only linear in the amount of test data, one can easily create a predictor that completely overfits, having empirical error on the test data that is artificially small [5, 12]. Even without such intentional overfitting, sequential querying can lead to unintentional adaptation since analysts are biased toward tweaks that lead to improved test errors.

If the queries were non-adaptive, i.e. the sequence of predictors is not influenced by previous test results, then we could handle a much larger number of queries before overfitting—a number exponential in the size of the dataset. Nevertheless, the test set will eventually be “used up” and estimates of the test error (specifically those of the best performers) might be over-optimistic.

A similar situation arises in other contexts such as validating potential scientific discoveries. One can evaluate potential discoveries using set aside validation data, but if analyses are refined adaptively based on the results, one may again overfit the validation data and arrive at false discoveries [14, 17]. One way to ensure the validity of answers in the face of adaptive querying is to collect all queries before giving any answers, and answer them all at once, e.g. at the end of a competition. However, analysts typically want more immediate feedback, both for ML challenges and in scientific research. Additionally, if we want to answer more queries later, ensuring statistical validity would require collecting a whole new dataset. This might be unnecessarily expensive if few or none of the queries are in fact adaptive. It also raises the question of who should bear the cost of collecting new data.

Alternatively, we could try to limit the number or frequency of queries from each user, forbid adaptive querying, or assume users work independently of each other, remaining oblivious to other users’ queries and answers. However, it is nearly impossible to enforce such restrictions. Determined users

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.’

can avoid querying restrictions by creating spurious user accounts and working in groups; there is no feasible way to check if queries are chosen adaptively; and information can leak between analysts, intentionally or not, e.g. through explicit collaboration or published results.

In this paper, we address the fundamental challenge of providing statistically valid answers to an arbitrarily long sequence of potentially adaptive queries. We assume that it is possible to collect additional samples from the same data distribution at a fixed cost per sample. To pay for new samples, users of the database will be charged for their queries. We propose a mechanism, EVERLASTING VALIDATION, that guarantees “everlasting” statistical validity and maintains the following properties:

Validity Without any assumptions about the users, and even with arbitrary adaptivity, with high probability, all answers ever returned by the database are accurate.

Self-Sustainability The database collects enough revenue to purchase as many new samples as necessary in perpetuity, and can answer an *unlimited* number of queries.

Cost for Non-Adaptive Users With high probability, a user making M non-adaptive queries will pay at most $O(\log M)$, so the average cost per query decreases as $O(1/M)$.

Cost for Autonomous Users With high probability, a user (or group of users) making M potentially adaptive queries that depend on each other arbitrarily, but not on any queries made by others, will pay at most $O(pM)$, so the average cost per query decreases as $O(1/pM)$.

We emphasize that the database mechanism needs no notion of “user” or “account” when answering the queries; it does not need to know which “user” made which query; and most of all, it does not need to know whether a query was made adaptively or not. Rather, the cost guarantees hold for any collection of queries that are either non-adaptive or autonomous in the sense described above—a “user” could thus refer to a single individual, or if an analyst uses answers from another person’s queries, we can consider them together as an “autonomous user” and get cost guarantees based on their combined number of queries. The database’s cost guarantees are nearly optimal; the cost to non-adaptive users and the cost to autonomous users cannot be improved (beyond log-factors) while still maintaining validity and sustainability (Section 5).

As is indicated by the guarantees above, using the mechanism adaptively may be far more expensive than using it non-adaptively. We view this as a positive feature. Although we cannot enforce nonadaptivity, and it is sometimes unreasonable to expect that analysts are entirely non-adaptive, we intend the mechanism to be used for *validation*. That is, analysts should do their discovery, training, tuning, development, and adaptive data analysis on unrestricted “training” or “discovery” datasets, and only use the protected database when they wish to receive a stamp of approval on their model, predictor, or discovery. Instead of trying to police or forbid adaptivity, we discourage it with pricing, but in a way that is essentially guaranteed not to affect non-adaptive users. Further, users will need to pay a high price only when their queries explicitly cause overfitting, so only adaptivity that is harmful to statistical validity will be penalized.

Relationship to prior work Our work is inspired by a number of mechanisms for dealing with potentially adaptive queries that have been proposed and analyzed using techniques from differential privacy and information theory. These mechanisms handle only a pre-determined number of queries using a fixed dataset. We use techniques developed in this literature, in particular addition of noise to ensure that a quadratically larger number of adaptive queries can be answered in the worst case [6, 10]. Our main innovations over this prior work are the self-sustaining nature of the database, as opposed to handling only a pre-determined number of queries of each type, and also the per-query pricing scheme that places the cost burden on the

adaptive users. To ensure that the cost burden on non-adaptive users does not grow by more than a constant factor, we need to adapt existing algorithms.

LADDER [5] and SHAKYLADDER [16] are mechanisms tailored to maintaining a ML competition leaderboard. These algorithms reveal the answer to a user’s query for the error of their model only if it is significantly lower than the error of the previous best submission from the user. While these mechanisms can handle an exponential number of arbitrarily adaptive submissions, each user will receive answers to a relatively small number of queries. Our setting is more suitable for the case where we want to validate the errors of all submissions or for scientific discovery where there is more than one discovery to be made.

A separate line of work in the statistics literature on “Quality Preserving Databases” (Aharoni and Rosset [2] and references therein) has suggested schemes for databases that maintain everlasting validity, while charging for use. The fundamental difference from our work is that these schemes do not account for adaptivity and thus are limited to non-adaptive querying. A second difference is that they focus on hypothesis testing for scientific discovery, with pricing schemes that depend on considerations of statistical power, which are not part of our framework. We further compare with existing methods at the end of Section 4.

2 Model formulation

We consider a setting in which a database curator has access to samples from some unknown distribution D over a sample space X . Multiple analysts submit a sequence of statistical queries $q_i : X \rightarrow [0,1]$, the database responds with answers $a_i \in \mathbb{R}$, and the goal is to ensure that with high probability, all answers satisfy $|a_i - \mathbb{E}_{x \leftarrow D}[q_i(x)]| \leq \epsilon$ for some fixed accuracy parameter ϵ . In a prediction validation application, each query would measure the expected loss of a particular model, while in scientific applications a single query might measure the value of some phenomenon of interest, or compare it to a “null” reference. We denote \mathcal{Q} the set of all possible queries, i.e. measurable

(function) shorthand for the empirical mean of desired answer) for each query. Given a data sample $q : X \rightarrow [0,1]$, and use the shorthand q on S . $\mathbb{E}[qS] = \mathbb{E}_{x \leftarrow D} [q(x)]$, we use $\mathbb{E}_D[q(x)]$ to denote the mean values $[q] = \mathbb{E}_D[q(x)]$ as

In our framework, the database can, at any time, acquire new samples from D at some fixed cost per sample, e.g. by running more experiments or paying workers to label more data. To answer a given query, the database can use the samples it has already purchased in any way it chooses, and the database is allowed to charge analysts for their queries in order to purchase additional samples. The price p_i of query q_i may be determined by the database after it receives query q_i , allowing the database to charge more for queries that force it to collect more data.

We do not assume the queries are chosen in advance, and instead allow the sequence of queries to depend adaptively on past answers. More formally, we define a “querying rule” $R_i : (\mathcal{Q}, \mathbb{R})^{i-1} \rightarrow \mathcal{Q}$ as a randomized mapping from the history of all previously made queries and their answers and prices to the statistical query to be made next:

$$q_i = R_i((q_1, a_1, p_1), (q_2, a_2, p_2), \dots, (q_{i-1}, a_{i-1}, p_{i-1})).$$

The interaction of users with the database can then be modeled as a sequence of querying rules

rules together define a joint distribution over queries, answers, and prices. Results will hold for any data distribution and any querying sequence, with high probability over $\{R_i\}_{i \in [N]}$. The combination of the data distribution, database mechanism, and sequence of querying $\{Q_i, A_i, P_i\}_{i \in [N]}$. All our

$\{Q_i, A_i, P_i\}_{i \in [N]}$.

We think of the query sequence as representing a combination of queries from multiple users, but the database itself is unaware of the identity or behavior of the users. Our validity guarantees do not assume any particular user structure, nor any constraints on the interactions of the different users. Thus, the guarantees are always valid regardless of what a “user” means, how “users” are allowed to collaborate, how many “users” there are, or how many queries each “user” makes—the guarantees simply hold for any (arbitrarily adaptive) querying sequence.

However, our cost guarantees will, and must, refer to analysts (or perhaps groups of analysts) behaving in specific ways. In particular, we define aof queries which do not depend on *any* of the history, i.e. non-adaptive user R_{u_j} as a subsequence is a fixed (pre-determined) distribution $\{u_j\}_{j \in [M]}$ consisting over queries, so Q_{u_j} is independent of all of the history. We further define an autonomous user of the database as a subsequence *the subsequence*, i.e. $\{u_j\}_{j \in [M]}$ of the querying rules that depend only on the history *within*

$$R_{u_j}(q_1, a_1, p_1), \dots, (q_{(u_j)}, a_{(u_j)}, p_{(u_j)}) =$$

$$R_{u_j}(q_{u_j}, a_{u_j}, p_{u_j}), \dots, (q_{u_{(j-1)}}, a_{u_{(j-1)}}, p_{u_{(j-1)}}) . \text{ That is, } Q_{u_j} \text{ is}$$

independent of the overall past history given the past history pertaining to the autonomous user.

The “cost to a user” is the total price paid for queries in the subsequence $\{u_j\}$:

■

3 VALIDATIONROUND

Our mechanism for providing “everlasting” validity guarantees is based on a query answering mechanism which we call non-adaptive and at least $\text{VALIDATIONROUND}(\epsilon, n^2 R)$ adaptive statistical queries (and potentially many more).. It uses n samples from D in order to answer $\exp(\epsilon)(n)$

Our analysis is based on ideas developed in the context of adaptive data analysis [10] and relies on techniques from differential privacy [9]. Differential privacy is a strong stability property of randomized algorithms that operate on a dataset. Composition properties of differential privacy imply that this form of stability holds even when the same dataset is used by multiple algorithms that can depend on the outputs of preceding algorithms. Most importantly, differential privacy implies generalization with high probability [4, 10].

VALIDATIONROUND splits its data into two sets S and T . Upon receiving each query, it first checks whether the answers on these datasets approximately agree. If so, the query has almost certainly not overfit to the data, and the algorithm simply returns the empirical mean of the query on S plus additional random noise. We show that the addition of noise ensures that the algorithm, as a function from the data sample S to an answer, satisfies differential privacy. This can be leveraged to show that any query which depends on a limited number of previous queries will have an empirical mean on S that is close to the true expectation. This ensures that VALIDATIONROUND can accurately answer a large number of queries, while allowing some (unknown) subset of the queries to be adaptive.

VALIDATIONROUND uses truncated Gaussian noise $\leftarrow \leftarrow N(0, \sigma^2, [-\beta, \beta])$, i.e. Gaussian noise $Z \leftarrow N(0, \sigma^2)$

conditioned on the event $|Z| \leq \beta$. Its density $f_{\text{trunc}}(x) = \exp(-x^2/(2\sigma^2)) / \exp(-\beta^2/(2\sigma^2))$.

[REDACTED]

1: [REDACTED] $\downarrow 4 \downarrow 8 \cdot \frac{32 \ln(8)}{\sigma^2}$ [REDACTED]

2: for each query q_1, q_2, \dots do

3: [REDACTED] if then

4: [REDACTED] Draw

truncated Gaussian [REDACTED] 5:6:

Output: $a_i = \mathbb{E}_S[q_i] + \leftarrow i$ else

7: Halt ($\mathcal{H} = i$)

Here, \mathcal{H} is the index of the query that causes the algorithm to halt. If $\mathcal{H} \leq I(\mathcal{X}, n)$, the maximum allowed number of answers, we say that VALIDATIONROUND halted “prematurely.” The following three lemmas characterize the behavior of VALIDATIONROUND.

Lemma 1. For any \mathcal{X} , σ , and n , for any sequence of querying rules (with arbitrary adaptivity) and any probability distribution \mathcal{D} , the answers provided by $\text{VALIDATIONROUND}(\mathcal{X}, n, S, T)$ satisfy

$$\mathbb{P} \left[\sum_{i=1}^{\mathcal{H}} |a_i - \mathbb{E}_S[q_i]| \geq \epsilon \right] \leq \exp(-\epsilon^2 / (8 \sigma^2 I(\mathcal{X}, n)))$$

where the probability is taken over the randomness in the draw of datasets S and T from \mathcal{D}^n , the querying rules, and VALIDATIONROUND.

Lemma 2. For any \mathcal{X} , σ , and n , any sequence of querying rules, and any non-adaptive user $\{u_j\}_{j \in [M]}$

interacting with $\text{VALIDATIONROUND}(\mathcal{X}, n, S, T)$, $\mathbb{P} \left[\sum_{i=1}^{\mathcal{H}} |a_i - \mathbb{E}_S[q_i]| \geq \epsilon \right] \leq \exp(-\epsilon^2 / (8 \sigma^2 \sum_{j \in [M]} I(\mathcal{X}, n)))$.

Lemma 3. For any \mathcal{X} , σ , and n , any sequence of querying rules, and any autonomous user $\{u_j\}_{j \in [M]}$

interacting with $\text{VALIDATIONROUND}(\mathcal{X}, n, S, T)$, if [REDACTED] and [REDACTED] then

$\mathbb{P} \left[\sum_{i=1}^{\mathcal{H}} |a_i - \mathbb{E}_S[q_i]| \geq \epsilon \right] \leq \exp(-\epsilon^2 / (8 \sigma^2 \sum_{j \in [M]} I(\mathcal{X}, n)))$.

Lemma 1 indicates that all returned answers are accurate with high probability, regardless of adaptivity. The proof involves showing that must be accurate since probability, non-adaptive queries never cause a premature halt, which is a simple application of $|\mathbb{E}_S[q_i] - \mathbb{E}_T[q_i]|$ [q is close to i]

and $|\leftarrow E|$ [are small. Lemma 2 indicates that with high q_i] for each query, so any query that is answered

$E\tau$

Hoeffding’s inequality. Finally, Lemma 3 shows that with high probability, an autonomous user who makes $O(\tilde{n}^2)$ queries will not cause a premature halt. This requires showing that $E s[q_i]$ is close to $E[q_i]$ despite the potential adaptivity.

The proof of Lemma 3 uses existing results from adaptive data analysis together with a simple argument that noise truncation does not significantly affect the results. For reference, the results we cite are included in Appendix E. While using Gaussian noise to answer queries is mentioned in other work, we are not aware of an explicit analysis, so we analyze the method here. To simplify parts of the derivation, we rely on the notion of concentrated differential privacy, which is particularly well suited for analysis of composition with Gaussian noise addition [6]. Lemmas 1-3 are proven in Appendix A.

4 EVERLASTINGVALIDATION and pricing

VALIDATIONROUND uses a fixed number, n , of samples and with high probability returns accurate answers for at least $\exp(-\tilde{n})$ non-adaptive queries and \tilde{n}^2 adaptive queries. In order to handle infinitely many queries, we chain together multiple instances of VALIDATIONROUND. We start with an initial dataset, answer queries using VALIDATIONROUND using that data until it halts. At this point, we buy more data and repeat. The used-up data can be released to the public as a “training set,” which can be used with no restriction without affecting any guarantees.

0
2
 $0 = 2, t = 0, i = 0$

```

3: Buy datasets  $S_0, T_0 \leftarrow D^{N_0}$ 
4: loop
5:   Pass  $q_i$  to VALIDATIONROUND( $\tilde{n}, t, N_t, S_t, T_t$ )
6:   if VALIDATIONROUND does not halt then 7:
       Output:  $a_i$ 
8:   Charge, [REDACTED] move on to  $i = i + 1$ 
9:   else
10:    Charge  $6N_t$  minus current capital
11:     $N_{t+1} = 3N_t, t+1 = \lfloor 2t \rfloor, t = t + 1$ 
12:    Buy datasets  $S_t, T_t \leftarrow D^{N_t}$ 
13:    Restart loop with same  $i$ 

```

The key ingredient is a pricing system with which we can always afford new data when an instance of VALIDATIONROUND halts. Our method has two price types: a low price, which is charged for all queries and decreases like $1/i$; and a high price, which is charged for any query that causes an instance of VALIDATIONROUND to halt prematurely, which may grow with the size of the current dataset. EVERLASTINGVALIDATION(\tilde{n}) guarantees the following:

Theorem 1 (Validity). *For any sequence of querying rules (with arbitrary adaptivity), EVERLASTINGVALIDATION will provide answers such that*

Proof. Consider the sequence of query rules that are answered by the t^{th} instantiation of the VALIDATIONROUND mechanism. By Lemma 1, for any sequence of querying rules, with probability

☐☐

The proof of Lemma 4 involves a straightforward computation. We find an upper bound, B_T , on the number of queries made before round T begins and then lower bound the revenue collected in round \square . We defer the details to Appendix B. Theorem 3 (Cost for non-adaptive users). *For any sequence of querying rules and any non-adaptive user indexed by $\{u_j\}_{j \in [M]}$, the cost to the user satisfies*

☐

(1)

15 JULY 2004

more than $6N_t$ price, then with probability at least $1 - t^{-2}$, inequality (1)

□

Relationship to prior work on adaptive data analysis We handle adaptivity using ideas developed in recent work on adaptive data analysis. In this line of work, all queries are typically assumed to be adaptively chosen and the overall number of queries known in advance. For completeness, we briefly describe several algorithms that have been developed in this context and compare them with our algorithm. Dwork et al. [10] analyze an algorithm that adds Laplace or Gaussian noise to the empirical mean in order to answer M adaptive queries using $O(\sqrt{pM})$ samples—a method that forms the basis of `VALIDATIONROUND`. However, adding untruncated Laplace or Gaussian noise to exponentially many non-adaptive queries would likely cause large errors when the variance is large enough to ensure that the sample mean is accurate. We use truncated Gaussian noise instead and show that it does not substantially affect the analysis for autonomous queries.

`THRESHOLDOUT` [11] answers verification queries in which the user submits both a query and an estimate of the answer. The algorithm uses $n = O(\sqrt{pM} \cdot \log I)$ samples to answer I queries of which at most M estimates are far from correct. Similar to our use of the second dataset T , this algorithm can be used to detect overfitting and answer adaptive queries (this is the basis of the `EFFECTIVEROUNDS` algorithm [10]). However, in our application this algorithm would have sample complexity of $n = O(\sqrt{pM} \cdot \log I)$, for M autonomous queries in T total queries. Consequently, direct use of this mechanism would result in a pricing for non-adaptive users that depends on the number of queries by autonomous users. This is in contrast to $n = O(\sqrt{pM} + \log T)$ samples that suffice for `VALIDATIONROUND`, where the improvement relies on our definition of autonomy and truncation of the noise variables.

5 Optimality

One might ask if it is possible to devise a mechanism with similar properties but lower costs. We argue that the prices set by `EVERLASTINGVALIDATION` are near optimal. The total cost to a non-adaptive user who makes M queries is $O(\log M / \epsilon^2)$. Even if we knew in advance that we would receive only M non-adaptive queries, we would still need $\Omega(\log M / \epsilon^2)$ samples to answer all of them accurately with high probability. Thus, our price for non-adaptive queries is optimal up to constant factors.

It is also known that answering a sequence of M adaptively chosen queries with accuracy ϵ requires

$\Omega(\sqrt{pM} / \epsilon)$ samples [15, 19]. Hence, the cost to a possibly adaptive autonomous user is nearly optimal in its dependence on M (up to log factors). One natural concern is that our guarantee in this case is only for the amortized (or total) cost, and not on the cost of each individual query. Indeed, although the *average* cost of adaptive queries decreases as $O(1/\sqrt{pM})$, the *maximal* cost of a single query might increase as $O(\sqrt{pM})$. A natural question is whether the maximum price can be reduced, to spread the high price over more queries.

Finally, an individual who queries our mechanism with M entirely non-adaptive queries will only pay $\log M$ in the worst case; generally, they will benefit from the economies of scale associated with collecting more and more data. For instance, if there are K users each making M non-adaptive queries, then the total cost of all KM queries will be $\log KM$ so the average cost to each user is only $\log(KM)/K \leq \log M$.

6 An Alternative Approach: EVERLASTINGTO

The EVERLASTINGVALIDATION mechanism provides cost guarantees that are, in certain ways, nearly optimal. The two main shortcomings are that (1) the price is guaranteed only for non-adaptive or autonomous users—not arbitrary adaptive ones and (2) the cost of an individual adaptive query cannot be upper bounded. One might also ask if inventing VALIDATIONROUND was necessary in the first place. Another mechanism, THRESHOLDOUT [11], is already well-suited to the setting of mixed adaptive and non-adaptive queries and it gives accuracy guarantees for quadratically many arbitrary adaptive queries or exponentially many non-adaptive queries. Perhaps using THRESHOLDOUT instead would be better? We will now describe an alternative mechanism, EVERLASTINGTO, which allows us to provide price guarantees for individual queries, including arbitrarily adaptive ones, but with an exponential increase in the cost for both non-adaptive and adaptive queries.

The EVERLASTINGTO mechanism is very similar to EVERLASTINGVALIDATION, except it uses THRESHOLDOUT in the place of VALIDATIONROUND. In each round, the algorithm determines an overfitting budget, B_t , and a maximum number of queries, M_t , as a function of the tradeoff parameter p . It then answers queries using THRESHOLDOUT, charging a high price $2N_{t+1}/B_t$ for queries that fail the overfitting check, and charging a low price $2N_{t+1}/M_t$ for all of the other queries. Once THRESHOLDOUT cannot answer more queries, the mechanism buys more data, reinitializes THRESHOLDOUT, and continues as before.

We analyze EVERLASTINGTO in Appendix D. Theorems 6-9 closely parallel the guarantees of ValidityEVERLASTING: with high probability, for any sequence of querying rules, all answers provided by VALIDATION and establish the following for any $\epsilon \in (0, 1)$ and any $p \in (0, 1)$:

EVERLASTINGTO are ϵ -accurate. Sustainability: EVERLASTINGTO charges high enough prices to be able to afford new samples as needed, excluding the initial budget. Cost: with high probability, any

Algorithm 3 EVERLASTINGTO(ϵ, p)

```

1:  $n = n(\epsilon, p)$ 
2:
3: for  $t = 0, 1, \dots$  do
4:   Purchase datasets  $S_t, T_t \leftarrow D^{N_t}$  and initialize THRESHOLDOUT( $S_t, T_t, B_t$ ,
5:     while THRESHOLDOUT( $S_t, T_t, B_t, \epsilon$ ) has not halted do
6:       Accept query  $q$ 
7:        $(a, o) = \text{THRESHOLDOUT}(S_t, T_t, B_t, \epsilon)(q)$ 
8:       Output:  $a$ 
9:       if  $o = ?$  then
10:        Charge:  $N$ 
11:       else
12:        Charge:

```

M non-adaptive queries and any B adaptive queries cost at most (ignoring the dependence on ϵ, p).

Unlike EVERLASTINGVALIDATION, which prioritized charging as little as possible for non-adaptive queries, EVERLASTINGTO increases the $O(\log M)$ cost to $O(\text{polylog} M)$ in order to bound the price of arbitrary adaptive queries. The parameter p allows the database manager to control the tradeoff; for p near zero, the cost of B adaptive queries is roughly the optimal $O(pB)$, but non-adaptive queries are extremely expensive. On the other side, for p near $2/3$, the cost of adaptive queries becomes very high, but the cost of non-adaptive queries is relatively small, although it does not approach optimality. Further details of the mechanism are contained in Appendix D. We also provide a tighter analysis of the THRESHOLDOUT algorithm which guarantees accurate answers using

a substantially smaller amount of data in Appendix C. This analysis allows us to reduce the exponent in EVERLASTINGTO's cost guarantee for non-adaptive queries.

7 Potential applications

In the ML challenge scenario, validation results are often displayed on a scoreboard. Although it is often assumed that scoreboards cannot be used for extensive adaptation, it appears that such adaptations have played roles in determining the outcome of various well known competitions, including the Netflix challenge, where the final test set performance was significantly worse than performance on the leaderboard data set. EVERLASTINGVALIDATION would guarantee that test errors returned by the validation database are accurate, regardless of adaptation, collusion, the number of queries made by each user, or other intentional or unintentional dependencies. We do charge a price per-validation, but as long as users are non-adaptive, the price is very small. Adaptive users, on the other hand, pay what is required in order to ensure validity (which could be a lot). Nevertheless, even if a wealthy user could afford paying the higher cost of adaptive queries, she would still not be able to “cheat” and overfit the scoreboard set, and a poor user could still afford the quickly diminishing costs of validating non-adaptive queries.

Another feature of our mechanism is that once a round t is over, we can safely release the datasets S_t and T_t to the public as unrestricted training data. This way, poor analysts also benefit from adaptive queries made by others, as all data is eventually released, and at any given time, a substantial fraction of all the data ever collected is public. Also, the ratio of public data to validation data can easily be adjusted by slightly amending the pricing.

In the context of scientific discovery, one use case is very similar to the ML competition. Scientists can search for interesting phenomena using unprotected data, and then re-evaluate “interesting” discoveries with the database mechanism in order to get an accurate and almost-unbiased estimate of the true value. This could be useful, for example, in building prediction models for scientific phenomena such as genetic risk of disease, which often involve complex modeling [7].

However, most scientific research is done in the context of hypothesis testing rather than estimation.

Declarations of discoveries like the Higgs boson [1] and genetic associations of disease [8] are based on performing a potentially large number of hypothesis tests and identifying statistically significant discoveries while controlling for multiplicity. Because of the complexity of the discovery process, it is often quite difficult to properly control for all potential tests, causing many difficulties, the most well known of which is the problem of publication bias (cf. “Why Most Published Research Findings are False” [17]). An alternative, approach that has gained popularity in recent years, is requiring replication of any declared discoveries on new and independent data [3]. Because the new data is used only for replication, it is much easier to control multiplicity and false discovery concerns.

Our everlasting database can be useful in both the discovery and replication phases. We now briefly explain how its validity guarantees can be used for multiplicity control in testing. Assume we have a collection of hypothesis tests on functionals of D_{q_i} . Setting with null hypotheses: $\epsilon = 1/2$, Theorem (1) guarantees: $H_{0i} : E[q_i] = e_{0i}$. We

P employ our scheme to obtain estimates A_i of $E[$

policy i $PHreject$ if $_{0i} [\max |A| A_i - e_{0i}|_{0i} > \alpha] > \alpha$ makes no false rejections with probability at

least α $\alpha - \epsilon$, meaning that for any combination of true nulls, the rejection ϵ . This is easily used in

the replication phase, where an ϵ , thus controlling the family-wise error rate (FWER) at level

entire community (say, type-I diabetes researchers) could share a single replication server using the everlasting database scheme in order to to guarantee validity. It could also be used in the

discovery phase for analyses that can be described through a set of measurements and tests of the form above. **8 Conclusion and extensions**

Our primary contribution is in designing a database mechanism that brings together two important properties that have not been previously combined: everlasting validity and robustness to adaptivity. Furthermore, we do so in an asymptotically efficient manner that guarantees that non-adaptive queries are inexpensive with high probability, and that the potentially high cost of handling adaptivity only falls upon truly adaptive users. Currently, there are large constants in the cost guarantees, but these are pessimistic and can likely be reduced with a tighter analysis and more refined pricing scheme. We believe that with some improvements, our scheme can form the basis of practical implementations for use in ML competitions and scientific discovery. Also, our cost guarantees themselves are worst-case and only guarantee a low price to entirely non-adaptive users. It would be useful to investigate experimentally how much users would actually end up being charged under “typical use,” especially users who are only “slightly adaptive.” However, there is no established framework for understanding what would constitute “typical” or “slightly adaptive” usage of a statistical query answering mechanism, so more work is needed before such experiments would be insightful.

Our mechanism can be improved in several ways. It only provides answers at a fixed, additive ϵ , and only answers statistical queries, however these issues have been already addressed in the adaptive data analysis literature. E.g. arbitrary low-sensitivity queries can be handled without any modification to the algorithm, and arbitrary real-valued queries can be answered with the error proportional to their standard deviation (instead of $1/pn$ as in our analysis) [13]. These approaches can be combined with our algorithms but we restrict our attention to the basic case since our focus is different.

Finally, one potentially objectionable element of our approach is that it discards samples at the end of each round (although these samples are not wasted since they become part of the public dataset). An alternative approach is to add the new samples to the dataset as they can be purchased. While this might be a more practical approach, existing analysis techniques that are based on differential privacy do not appear to suffice for dealing with such mechanisms. Developing more flexible analysis techniques for this purpose is another natural direction for future work.

Acknowledgements BW is supported the NSF Graduate Research Fellowship under award 1754881.

References

- [1] Georges Aad, T Abajyan, B Abbott, J Abdallah, S Abdel Khalek, AA Abdelalim, O Abdinov, R Aben, B Abi, M Abolins, et al. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1–29, 2012.
- [2] Ehud Aharoni and Saharon Rosset. Generalized alpha-investing: Definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B*, 76(4): 771–794, 2014.
- [3] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452, 2016.
- [4] Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, pages 1046–1059, 2016.
- [5] Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, pages 1006–1014, 2015.
- [6] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

- [7] Nilanjan Chatterjee, Jianxin Shi, and Montserrat Garc'ia-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17 (7):392, 2016.
- [8] Nick Craddock, Matthew E Hurles, Niall Cardin, Richard D Pearson, Vincent Plagnol, Samuel Robson, Damjan Vukcevic, Chris Barnes, Donald F Conrad, Eleni Giannoulidou, et al. Genomewide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713, 2010.
- [9] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.
- [10] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. *CoRR*, abs/1411.2664, 2014. Extended abstract in STOC 2015.
- [11] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015.
- [12] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349 (6248):636–638, 2015. doi: 10.1126/science.aaa9375. URL <http://www.sciencemag.org/content/349/6248/636.abstract>.
- [13] Vitaly Feldman and Thomas Steinke. Generalization for adaptively-chosen estimators via stable median. In *Conference on Learning Theory (COLT)*, 2017.
- [14] Andrew Gelman and Eric Loken. The statistical crisis in science. *The American Statistician*, 102(6):460, 2014.
- [15] M. Hardt and J. Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS*, pages 454–463, 2014.
- [16] Moritz Hardt. Climbing a shaky ladder: Better adaptive risk estimation. *CoRR*, abs/1706.02733, 2017. URL <http://arxiv.org/abs/1706.02733>.
- [17] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.
- [18] Kobbi Nissim and Uri Stemmer. On the generalization properties of differential privacy. *CoRR*, abs/1504.05800, 2015.
- [19] Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *COLT*, pages 1588–1628, 2015. URL <http://jmlr.org/proceedings/papers/v40/Steinke15.html>.