The Everlasting Database: Statistical Validity at a Fair Price

Blake Woodworth*[1], Vitaly Feldman[2], Saharon Rosset[3] and Nathan Srebro[1]

[1]Toyota Technological Institute at Chicago
[2]Google Brain
[3]School of Mathematical Sciences, Tel Aviv University

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 $O(\sqrt{M})$.

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.

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

If the queries were non-adaptive, i.e. the sequence of predictors submitted is not influenced by previous test results, then we could handle a much larger number of queries before overfitting, namely a number of queries exponential in the size of the dataset. Nevertheless, eventually the test set could 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 scientific discoveries based on data analysis. We can set aside validation data and use it to evaluate potential discoveries, but if analyses are refined adaptively based on the results, we may again overfit the validation data and arrive at false discoveries [16, 13].

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

*This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1754881.

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 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 ensuring statistical validity of 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 some fixed cost per sample. To pay for additional samples, the users of the database will be required to pay for their queries. We propose a mechanism, EverlastingValidation, that guarantees "everlasting" statistical validity and maintains the following properties:

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.

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

Cost for Non-Adaptive Users A user making M non-adaptive queries is guaranteed that, with high probability, her total cost will be at most $O(\log M)$, so the average cost per query decreases as $\tilde{O}(1/M)$.

Cost for Autonomous Users 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, is guaranteed that, with high probability, her cost will be at most $\tilde{O}(\sqrt{M})$, so the average cost per query decreases as $\tilde{O}(1/\sqrt{M})$.

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 is far more expensive than using it non-adaptively. We view this as a positive feature. Although we cannot enforce non-adaptivity, and it is sometimes unreasonable to expect that analysts are entirely non-adaptive, the intended use is 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. Therefore, 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 a pre-determined number of queries using a fixed dataset. We rely on 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 [11]. Our main innovation over this prior work is 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 [15] are mechanisms tailored to maintaining a ML competition leaderboard. These algorithms reveal the answer to a user's query (submission of a model) only if the error of the model 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 then one discovery to be made. At the same time, those mechanisms can easily be combined with our approach to making validity guarantees everlasting by collecting additional data and penalizing only adaptive users.

A separate line of work in the statistics literature on "Quality Preserving Databases" ([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 $\mathcal{D}$ over a sample space $\mathcal{X}$. Multiple analysts interact with the database by posing queries about $\mathcal{D}$. Analysts submit a sequence of statistical queries $q_i : \mathcal{X} \to [0,1]$, the database responds with answers $a_i \in \mathbb{R}$, and the goal is to ensure that with high probability, all answers satisfy $|a_i - \mathbb{E}_{x\sim\mathcal{D}}[q_i(x)]| \leq \tau$ for some fixed accuracy parameter τ. In a prediction validation application, each query measures the expected error or 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 functions $q : \mathcal{X} \to [0,1]$, and use the shorthand $\mathbb{E}[q] = \mathbb{E}_{x\sim\mathcal{D}}[q(x)]$ to denote the mean value (desired answer) for each query. Given a data sample $S \sim \mathcal{D}^n$, we use $\mathcal{E}_S[q] = \frac{1}{|S|}\sum_{x\in S} q(x)$ as shorthand for the empirical mean of q on S.

We consider a setting in which the database can, at any time, acquire new samples from $\mathcal{D}$ at some fixed cost per sample, e.g. by running more experiments, making more observations, 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, thus the database can charge more for the 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" $R_i : (\mathcal{Q}, \mathbb{R}, \mathbb{R})^{i-1} \mapsto \mathcal{Q}$ 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, i.e.:

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

We can then model the interaction of users with the database as a sequence of querying rules $\{R_i\}_{i\in\mathbb{N}}$. The combination of the data distribution, the database mechanism, and the sequence of querying rules together define a joint distribution over queries, answers, and prices $\{Q_i, A_i, P_i\}_{i\in\mathbb{N}}$. All our results will hold for any data distribution, any querying sequence, and with high probability over $\{Q_i, A_i, P_i\}_{i\in\mathbb{N}}$.

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 an **autonomous user** of the database as a subsequence $\{u_j\}_{j\in[M]}$ of the querying rules such that all querying rules within the subsequence depend only on the history *within the subsequence*, i.e.

$$R_{u_j}\left((q_1,a_1,p_1),\ldots,(q_{(u_j-1)},a_{(u_j-1)},p_{(u_j-1)})\right) \quad = \quad R_{u_j}\left((q_{u_1},a_{u_1},p_{u_1}),\ldots,(q_{u_{(j-1)}},a_{u_{(j-1)}},p_{u_{(j-1)}})\right).$$

That is, Q_{u_j} is independent of the overall past history given the past history pertaining to the autonomous user. We further define a **non-adaptive user** as a subsequence consisting of the queries which do not depend on *any* of the history, i.e. $R_{u_j}\left((q_1,a_1,p_1),\ldots,(q_{(u_j-1)},a_{(u_j-1)},p_{(u_j-1)})\right)$ is a fixed (pre-determined) distribution over queries, and so Q_{u_j} is independent of all of the history. The cost to a user (total price paid for queries) is $\sum_{j=1}^{M} p_{u_j}$.

3 ValidationRound

The mechanism that provides "everlasting" validity guarantees is based on a query answering mechanism which we call ValidationRound. It uses n samples from $\mathcal{D}$ in order to answer $\exp(\Omega(n))$ non-adaptive and $\tilde{\Omega}(n^2)$ adaptive statistical queries. The analysis of the adaptive case is based on ideas developed in the context of adaptive data analysis [11] 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 [11, 4].

ValidationRound splits its data into two sets S and T. Upon receiving each query, ValidationRound first checks whether the answers on these datasets approximately agree. If so, the query most likely has 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 $\xi \sim \mathcal{N}(0,\sigma^2,[-\gamma,\gamma])$, i.e. Gaussian noise $Z \sim \mathcal{N}(0,\sigma^2)$ conditioned on the event $|Z| \le \gamma$. Its density $f_\xi(x) \propto \exp\left(-\frac{x^2}{2\sigma^2}\right)\mathbb{1}_{|x|\le\gamma}$.

Algorithm 1 ValidationRound(τ,β,n,S,T)

Set $I(\tau,\beta,n) = \frac{\beta}{4}\exp\left(\frac{n\tau^2}{8}\right)$, $\sigma^2 = \frac{\tau^2}{32\ln(8n^2/\beta)}$
for each query $q_1, q_2, \ldots$ **do**
  if $|\mathcal{E}_S[q_i] - \mathcal{E}_T[q_i]| \le \frac{\tau}{2}$ **and** $i \le I(\tau,\beta,n)$ **then**
    Draw truncated Gaussian $\xi_i \sim \mathcal{N}(0,\sigma^2,[-\frac{\tau}{4},\frac{\tau}{4}])$
    Output: $a_i = \mathcal{E}_S[q_i] + \xi_i$
  else
    Halt $(\eta = i)$

Here, η is the index of the query that causes the algorithm to halt. If $\eta \le I(\tau,\beta,n)$, 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 τ, β, and n, for any sequence of querying rules (with arbitrary adaptivity) and any probability distribution $\mathcal{D}$,* ValidationRound(τ,β,n,S,T) *will return a sequence of answers before halting*

such that

$$\mathbb{P}\left[\forall_{i<\eta}\left|A_i - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q_i(x)]\right| \leq \tau\right] \geq 1 - \frac{\beta}{2},$$

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 τ, β, and n, any sequence of querying rules, and any non-adaptive user $\{u_j\}_{j\in[M]}$ interacting with* ValidationRound(τ, β, n, S, T), $\mathbb{P}\left[\eta \leq I(\tau,\beta,n) \wedge \eta \in \{u_j\}_{j\in[M]}\right] \leq \beta$.

Lemma 3. *For any τ, β, and n, any sequence of querying rules, and any possibly adaptive user $\{u_j\}_{j\in[M]}$ interacting with* ValidationRound(τ, β, n, S, T), *if $\sigma^2 = \frac{\tau^2}{32\ln(8n^2/\beta)}$ and $M \leq \frac{n^2\tau^4}{175760\ln^2(8n^2/\beta)}$ then*

$$\mathbb{P}\left[\eta \leq I(\tau,\beta,n) \wedge \eta \in \{u_j\}_{j\in[M]}\right] \leq \beta.$$

In other words, Lemmas 2 and 3 show that with high probability, a non-adaptive user will not cause ValidationRound to halt prematurely and an adaptive user will only do so if they have already made $\tilde{\Omega}(n^2)$ queries.

Proofs

Lemma 4. *For any τ, β, n, and any sequence of querying rules (with arbitrary adaptivity) interacting with* ValidationRound(τ, β, n, S, T)

$$\mathbb{P}\left[\forall_{i<\eta}\ \left|\mathcal{E}_T[Q_i] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q_i(x)]\right| \leq \frac{\tau}{4}\right] \geq 1 - \frac{\beta}{2}.$$

Proof. Consider any sequence of querying rules (with arbitrary adaptivity). The interaction between the query rules and ValidationRound(τ, β, n, S, T) together determines a joint distribution over statistical queries, answers, and prices $(Q_1, A_1, P_1), ..., (Q_{\eta-1}, A_{\eta-1}, P_{\eta-1})$.

Consider also the interaction of the same sequence of querying rules with an alternative algorithm, which always returns $\mathcal{E}_S[q_i] + \xi_i$ (i.e. it ignores the if-statement in ValidationRound). This generates an infinite sequence of queries, answers, and prices $(Q'_1, A'_1, P'_1), (Q'_2, A'_2, P'_2), ...$ Now, we retroactively check the condition in the if-statement for each of the queries to calculate what η should be, and take the length $\eta - 1$ prefix of the (Q'_i, A'_i, P'_i). This sequence has exactly the same distribution as the sequence generated by ValidationRound, and each Q'_i was chosen independently of T by construction. Since $Q'_i \sim Q_i$ has outputs bounded in $[0, 1]$, we can apply Hoeffding's inequality:

$$\mathbb{P}\left[\left|\mathcal{E}_T[Q_i] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q_i(x)]\right| > \frac{\tau}{4}\right] \leq 2\exp\left(-\frac{n\tau^2}{8}\right).$$

At most $I(\tau,\beta,n) = \frac{\beta}{4}\exp\left(\frac{n\tau^2}{8}\right)$ queries are answered by the mechanism, so a union bound completes the proof. □

Proof of Lemma 1. A query is not answered unless $|\mathcal{E}_S[q_i] - \mathcal{E}_T[q_i]| \leq \frac{\tau}{2}$, so $\forall i < \eta$

$$|a_i - \mathbb{E}[q_i]| \leq |\xi_i| + |\mathcal{E}_S[q_i] - \mathcal{E}_T[q_i]| + |\mathcal{E}_T[q_i] - \mathbb{E}[q_i]| \leq \tau/4 + \tau/2 + |\mathcal{E}_T[q_i] - \mathbb{E}[q_i]|.$$

By Lemma 4, with probabilty $1 - \frac{\beta}{2}$ the final term is at most $\tau/4$ simultaneously for all $i < \eta$. □

Proof of Lemma 2. Since the non-adaptive user's querying rules ignore all of the history, they are each chosen independently of S. By Hoeffding's inequality

$$\mathbb{P}\left[\left|\mathcal{E}_S\left[Q_{u_j}\right] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[Q_{u_j}(x)\right]\right| > \frac{\tau}{4}\right] \leq 2\exp\left(-\frac{n\tau^2}{8}\right)$$

and similarly for T. If both $\eta \leq I(\tau,\beta,n)$ and $\eta = u_j$, then the algorithm halted upon receiving query q_{u_j} because its empirical means on S and T were too dissimilar and *not* because it had already answered its maximum allotment of queries. Therefore,

$$\mathbb{P}\left[\eta \leq I(\tau,\beta,n) \wedge \eta = u_j\right] = \mathbb{P}\left[\left|\mathcal{E}_S\left[Q_{u_j}\right] - \mathcal{E}_T\left[Q_{u_j}\right]\right| > \frac{\tau}{2}\right] \leq 4\exp\left(-\frac{n\tau^2}{8}\right).$$

At most $I(\tau,\beta,n) = \frac{\beta}{4}\exp\left(\frac{n\tau^2}{8}\right)$ queries are answered by the mechanism, so a union bound completes the proof. □

Lemma 5. *For any τ, β, n, any sequence of query rules, and any possibly adaptive user $\{u_j\}_{j\in[M]}$, if $\sigma^2 = \frac{\tau^2}{32\ln(8n^2/\beta)}$ and $M \leq \frac{n^2\tau^4}{175760\ln^2(8n^2/\beta)}$ then*

$$\mathbb{P}\left[\forall_{j\in[M]} \ \left|\mathcal{E}_S\left[Q_{u_j}\right] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[Q_{u_j}(x)\right]\right| \leq \frac{\tau}{4}\right] \geq 1 - \frac{\beta}{2}.$$

Lemma 5 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 the supplementary materials. While using Gaussian noise to answer queries is mentioned in other work, we are not aware of explicitly stated bounds for it. Hence, 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 and Gaussian noise addition [6].

Proof of Lemma 5. Consider a slightly modified version of ValidationRound, where Gaussian noise $z_i \sim \mathcal{N}(0,\sigma^2)$ is added instead of truncated Gaussian noise ξ_i. Until this modified algorithm halts, all of the answers it provides are released according to the Gaussian mechanism on S, which satisfies $\frac{1}{2n^2\sigma^2}$-zCDP by Proposition 1.6 in [6]. We can view $Q_{u_j} = R_{u_j}((q_{u_1}, a_{u_1}, p_{u_1})..., (q_{u_{j-1}}, a_{u_{j-1}}, p_{u_{j-1}}))$ as an (at most) M-fold composition of $\frac{1}{2n^2\sigma^2}$-zCDP mechanisms, which satisfies $\frac{M}{2n^2\sigma^2}$-zCDP by Lemma 1.7 in [6]. Finally, Proposition 1.3 in [6] shows us how to convert this concentrated differential privacy guarantee to a regular differential privacy guarantee. In particular, q_{u_j} is generated under

$$\left(\frac{M}{2n^2\sigma^2} + 2\sqrt{\frac{M}{2n^2\sigma^2}\ln(1/\delta)},\ \delta\right)\text{-DP} \quad \forall\delta > 0.$$

Specifically, when σ^2, δ and M satisfy:

$$\begin{aligned}
\sigma^2 &= \frac{\tau^2}{32\ln(8n^2/\beta)} \\
\delta = \frac{\beta}{8n^2} &= \frac{\beta}{\frac{n^2\tau}{13\ln(104/\tau)}} \cdot \frac{\tau}{104\ln(104/\tau)} \\
M &\leq \frac{n^2\tau^4}{175760\ln^2(8n^2/\beta)}.
\end{aligned}$$

then $q_{,i_j}$ is generated by a $\left(\frac{\tau}{52},\delta\right)$-differentially private mechanism. Therefore, by Theorem 8 in [10] (cf. [17, 4])

$$\mathbb{P}\left[\left|\mathcal{E}_S\left[q_{u_j}\right] - \mathbb{E}\left[q_{u_j}\right]\right| > \frac{\tau}{4}\right] \leq \frac{\beta}{\frac{n^2\tau}{13\ln(104/\tau)}} \ll \frac{\beta}{4M}.$$

Furthermore, for $z_i \sim \mathcal{N}\left(0, \sigma^2\right)$ $\mathbb{P}\left[|z_i| \geq \tau/4\right] \leq \beta/(4n^2) \leq \beta/(4M)$. Therefore, the total variation distance between $\xi_{u_j} \sim \mathcal{N}\left(0, \sigma^2, [-\tau/4, \tau/4]\right)$ and $z_{u_j} \sim \mathcal{N}(0, \sigma^2)$ is $\Delta(\xi_{u_j}, z_{u_j}) = \mathbb{P}\left[z_{u_j} \notin [-\tau/4, \tau/4]\right] \leq \frac{\beta}{4M}$. Consider two random vectors Z and ξ, the first of which has independent $\mathcal{N}(0, \sigma^2)$ distributed coordinates, and the second of which has coordinates $\xi_{u_j} \sim \mathcal{N}\left(0, \sigma^2, [-\tau/4, \tau/4]\right)$ for $j \in [M]$ and $\xi_i = Z_i$ for all of the $i \notin \{u_j\}$. The total variation distance between these vectors is then at most $\Delta(\xi, Z) \leq M\Delta(\xi_{u_j}, z_{u_j}) \leq \frac{\beta}{4}$.

Now, for the given sequence of querying rules, S, and T, view ValidationRound as a function of the random noise which is added into the answers. Then $\Delta(\text{ValidationRound}(\xi), \text{ValidationRound}(Z)) \leq \Delta(\xi, Z) \leq \frac{\beta}{4}$ too. Above, we showed that with probability $1 - \beta/4$ the user's interaction with ValidationRound(Z) has the property that

$$\mathbb{P}\left[\exists_{j \in [M]} \ \left|\mathcal{E}_S\left[q_{u_j}\right] - \mathbb{E}\left[q_{u_j}\right]\right| > \frac{\tau}{4}\right] \leq \frac{\beta}{4}.$$

So their interaction with ValidationRound(ξ) satisfies

$$\mathbb{P}\left[\exists_{j \in [M]} \ \left|\mathcal{E}_S\left[q_{u_j}\right] - \mathbb{E}\left[q_{u_j}\right]\right| > \frac{\tau}{4}\right] \leq \frac{\beta}{2}.$$

Since this statement only depends on the indices of ξ in $\{u_j\}_{j \in [M]}$, we can replace all of the remaining indices with truncated Gaussians and maintain this property, which recovers ValidationRound. □

Proof of Lemma 3. Consider a query q_{u_j} made by the autonomous user. Lemma 4 guarantees that

$$\mathbb{P}\left[\forall_{j \in [M]} \ \left|\mathcal{E}_T\left[q_{u_j}\right] - \mathbb{E}\left[q_{u_j}\right]\right| \leq \frac{\tau}{4}\right] \geq 1 - \frac{\beta}{2}.$$

By Lemma 5, with the hypothesized σ^2 and M

$$\mathbb{P}\left[\forall_{j \in [M]} \ \left|\mathcal{E}_S\left[q_{u_j}\right] - \mathbb{E}\left[q_{u_j}\right]\right| \leq \frac{\tau}{4}\right] \geq 1 - \frac{\beta}{2}.$$

If both $\eta \leq I(\tau, \beta, n)$ and $\eta \in \{u_j\}_{j \in [M]}$, then the algorithm halted upon receiving a query q_{u_j} because its empirical means on S and T were too dissimilar and *not* because it had already answered its maximum allotment of queries:

$$\mathbb{P}\left[\eta \leq I(\tau, \beta, n) \wedge \eta \in \{u_j\}_{j \in [M]}\right] = \mathbb{P}\left[\exists_{j \in [M]} \ \left|\mathcal{E}_S\left[q_{u_j}\right] - \mathcal{E}_T\left[q_{u_j}\right]\right| > \frac{\tau}{2}\right] \leq \beta$$

□

4 EverlastingValidation and Pricing

ValidationRound uses a fixed number, n, of samples and with high probability returns accurate answers for at least $\exp(\Omega(n))$ non-adaptive queries and $\tilde{\Omega}(n^2)$ 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.

Algorithm 2 EVERLASTINGVALIDATION(τ, β)

1: Require initial budget $\Gamma = 36 \ln(8/\beta)/\tau^2$
2: $N_0 = \frac{\Gamma}{2}$, $\beta_0 = \frac{\beta}{2}$, $t = 0$, $i = 0$
3: Buy datasets $S_0, T_0 \sim \mathcal{D}^{N_0}$
4: **loop**
5: Pass q_i to VALIDATIONROUND$(\tau, \beta_t, N_t, S_t, T_t)$
6: **if** VALIDATIONROUND does not halt **then**
7: **Output:** answer
8: Charge $\frac{96}{\tau^2} \cdot \frac{1}{i}$, move on to $i = i + 1$
9: **else**
10: Charge $6N_t$ minus current capital
11: $N_{t+1} = 3N_t$, $\beta_{t+1} = \frac{1}{2}\beta_t$, $t = t + 1$
12: Buy datasets $S_t, T_t \sim \mathcal{D}^{N_t}$
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(τ, β) guarantees the following:

Theorem 1 (Sustainability)**.** *For any sequence of queries, the prices charged will be enough to pay for all samples needed by* EVERLASTINGVALIDATION *in perpetuity, excluding the initial budget of* $36 \ln(8/\beta)/\tau^2$*.*

Proof. When an instance of VALIDATIONROUND halts, we charge exactly enough for the next S_t, T_t (line 10). □

Theorem 2 (Validity)**.** *For any sequence of querying rules (with arbitrary adaptivity),* EVERLASTINGVALIDATION *will provide answers such that*

$$\mathbb{P}\left[\forall_{i \in \mathbb{N}} \left| A_i - \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[Q_i(x)] \right| \leq \tau \right] \geq 1 - \frac{\beta}{2}$$

Proof. Consider the sequence of query rules that are answered by the t^{th} instantiation of the VALIDATIONROUND mechanism. For the purposes of Lemma 1, it does not matter that these query rules may use the history of answers given before round t since that history did not depend on the current data S_t, T_t. With probability $1 - \frac{\beta_t}{2}$ all queries answered during round t are answered accurately (Lemma 1). By a union bound over all rounds, all answers are accurate with probabity at least $1 - \sum_{t=0}^{\infty} \beta_t/2 = 1 - \beta/2$. □

Lemma 6. *If* $N_0 \geq 18 \ln(2)/\tau^2$ *and* $I(\tau, \beta_t, N_t) = (\beta_t/4) \exp\left(N_t \tau^2/8\right)$ *queries are answered during round* t*, then at least* $6N_t$ *revenue is collected.*

Proof. The revenue collected in round t via the low price $\frac{96}{\tau^2 i}$ depends on how many queries are answered both in and before round t. The maximum number of queries answered in a round is $I_t = I(\tau, \beta_t, N_t) = (\beta_t/4) \exp\left(N_t \tau^2/8\right)$ (this is enforced by VALIDATIONROUND). Let B_T be the total number of queries made

before the beginning of round T, then

$$\begin{aligned}
B_T \le \sum_{t=0}^{T-1} I_t &= \sum_{t=0}^{T-1} \frac{\beta_t}{4} \exp\left(\frac{N_t \tau^2}{8}\right) \\
&= \frac{\beta_0}{4} \sum_{t=0}^{T-1} \exp\left(\frac{\tau^2}{8} 3^t N_0 - t \ln 2\right) \\
&\le \frac{\beta_0}{4} \exp\left(\sum_{t=0}^{T-1} \frac{\tau^2}{8} 3^t N_0 - t \ln 2\right) \\
&= \frac{\beta_0}{4} \exp\left(-\frac{T(T-1)}{2} \ln 2 + \frac{3^T - 1}{2} \frac{N_0 \tau^2}{8}\right) \\
&= \frac{\beta_0 2^{-T}}{4} \exp\left(\frac{-T^2 + 3T - \frac{N_0 \tau^2}{8 \ln 2}}{2} \ln 2 + \frac{N_T \tau^2}{16}\right) \\
&\le (\beta_T/4) \exp\left(N_T \tau^2/16\right).
\end{aligned}$$

The first inequality holds because every exponent in the sum is at least $\ln(2)$ by our choice of N_0 and for any $x, y \ge \ln 2$, $e^{x+y} \ge 2\max(e^x, e^y) \ge e^x + e^y$. The second inequality holds since $N_0 > \frac{18 \ln 2}{\tau^2}$ implies $-T^2 + 3T - N_0 \tau^2/(8 \ln 2) \le 0$. So, if I_T queries are answered during round T, the revenue collected is at least

$$\begin{aligned}
\sum_{i=1}^{I_T} \frac{96}{\tau^2 (B_T + i)} &= \frac{96}{\tau^2} \sum_{i=1}^{B_T + I_T} \frac{1}{i} - \sum_{i=1}^{B_T} \frac{1}{i} \\
&\ge \frac{96}{\tau^2} \left(\ln(B_T + I_T) - \ln(B_T)\right) \\
&= \frac{96}{\tau^2} \ln\left(1 + \frac{I_T}{B_T}\right) \\
&\ge \frac{96}{\tau^2} \ln\left(1 + \frac{(\beta_T/4) \exp\left(N_T \tau^2/8\right)}{(\beta_T/4) \exp\left(N_T \tau^2/16\right)}\right) \\
&\ge 6 N_T
\end{aligned}$$

□

Theorem 3 (Cost for non-adaptive users)**.** *For any sequence of querying rules and any non-adaptive autonomous user indexed by $\{u_j\}_{j \in [M]}$, the cost to the user satisfies*

$$\mathbb{P}\left[\sum_{j \in [M]} P_{u_j} \le \frac{96}{\tau^2} (1 + \ln(M))\right] \ge 1 - \beta.$$

Proof. By Lemma 6, if a round t ends after $I(\tau, \beta_t, N_t)$ queries are answered, then the total revenue collected from queries in that round is at least $6N_t$, so the "high price" at the end of the round is 0.

Consequently, a query q_{u_j} from the non-adaptive user will only cost more than the low price $\frac{96}{\tau^2 u_j}$ if it causes an instantiation of ValidationRound to halt prematurely. By Lemma 2 this occurs with probability at most β_t in round t. By a union bound over all rounds, this never happens with probability at least $1 - \sum_{t=0}^{\infty} \beta_t = 1 - \beta$. In this case, the cost to the user is

$$\sum_{j=1}^{M} p_{u_j} = \sum_{j=1}^{M} \frac{96}{\tau^2 u_j} \le \sum_{i=1}^{M} \frac{96}{\tau^2 i} \le \frac{96}{\tau^2} (1 + \ln(M)).$$

□

Theorem 4 (Cost for adaptive users)**.** *For any sequence of querying rules and any possibly adaptive autonomous user indexed by* $\{u_j\}_{j\in[M]}$*, the cost to the user satisfies*

$$\mathbb{P}\left[\sum_{j\in[M]} P_{u_j} \le c_0 \cdot \frac{\sqrt{M}\ln^2(M/\beta)}{\tau^2}\right] \ge 1-\beta,$$

where c_0 *is a fixed constant.*

Proof. Ideally, all M queries will cost the low price $\frac{96}{\tau^2 u_j}$, totalling at most $\frac{96}{\tau^2}(1+\ln(M))$. However, the adaptive user may cause rounds to end prematurely and pay (up to) $6N_t$. Nevertheless, by Lemma 3, with probability $1-\beta_t$ if one of the adaptive user's queries causes a round t to end prematurely, then the amount of data, N_t, and the number of the user's queries answered in that round, M_t, must satisfy

$$M_t \ge \frac{N_t^2\tau^4}{175760\ln^2(8N_t^2/\beta_t)}. \tag{1}$$

So, for a given M there is a largest t for which this is possible since $N_t = 3^t N_0$ and $\beta_t = 2^{-t}\beta_0$. In particular,

$$\frac{3^{2t}N_0^2\tau^4}{175760\ln(18^t\cdot 8N_0^2/\beta_0)} \le M$$

which implies $t_{\max} \le \frac{1}{2}\ln\left(24\sqrt{M}\ln(144N_0/\beta_0)\right)$. Let $\mathcal{T}$ be the set of rounds in which the adaptive user pays the high $6N_t$ price, then with probability at least $1-\sum_{t\in\mathcal{T}}\beta_t \ge 1-\beta$, inequality (1) holds for all $t\in\mathcal{T}$. In this case, the total cost to the adaptive user is no more than

$$\begin{aligned}\sum_{t\in\mathcal{T}} 6N_t &\le 6\sum_{t\in\mathcal{T}} \frac{420\sqrt{M_t}\ln\left(8M_t^2/\beta_t\right)}{\tau^2}\\ &\le t_{\max}\frac{2520\sqrt{M}\ln\left(8M^2/\beta_{t_{\max}}\right)}{\tau^2}\\ &\le \frac{1890\sqrt{M}\ln^2\left(16M^2/\beta\right)}{\tau^2}.\end{aligned}$$

□

Relationship to prior work on adaptive data analysis To handle adaptivity we rely on ideas developed in the context of the recent work on adaptive data analysis. In this line of work the entire sequence of queries is 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. [11] analyze an algorithm that adds Laplace or Gaussian noise to the empirical mean in order to answer M adaptive queries using $\tilde{O}(\sqrt{M})$ samples—a method that forms the basis of ValidationRound. However, adding (untruncated) Laplace or Gaussian noise to exponentially many non-adaptive queries would likely lead to large errors when used with variance necessary to ensure that adaptive queries are answered correctly. Thus we use truncated Gaussian noise instead and use the fact that it would not substantially affect the analysis for adaptive queries.

Thresholdout [10] answers verification queries in which the user submits both a query and an estimate of the answer. The algorithm uses $n = \tilde{O}(\sqrt{R}\cdot\log M)$ samples to answer M queries whenever at most of the estimates R are far from correct. This algorithm can be used to detect overfitting and answer adaptive queries in a manner similar to our use of the second dataset T (an example of such use can be found in EffectiveRounds algorithm [11]). However in our application this algorithm would have sample complexity of $n = \tilde{O}(\sqrt{M}\cdot\log T)$, where M is the number of queries that can be asked by an adaptive autonomous user and T is the total number of allowed queries. Direct use of this mechanism would result in a pricing scheme that penalizes non-adaptive users by an unbounded factor. This is in contrast to $n = \tilde{O}(\sqrt{M}+\log T)$ samples that suffice for ValidationRound. The improvement relies on our definition of autonomy and the 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 of a sequence of M queries is $O(\log M/\tau^2)$. Even if we knew in advance that we would receive only M non-adaptive queries, we would still need $\Omega(\log M/\tau^2)$ 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 τ requires $\tilde{\Omega}(\sqrt{M}/\tau)$ samples [14, 19]. Hence the cost to an autonomous, possibly adaptive user is nearly optimal (up to log factors) in its dependence on M. 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 $\tilde{O}(1/\sqrt{M})$ with the number of queries, the maximal cost of a single query might increase as $\sqrt{M}$. A natural question is whether the maximum price can be reduced, to spread high price over more queries.

6 Potential Applications

Two possible applications of our method are validating results in ML challenges, and validating scientific discoveries.

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, and the Baidu ImageNet "success" [18]. Our mechanism, on the other hand, guarantees 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 company could afford paying the higher cost of adaptive queries, it 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. Furthermore, wealthy companies can in any case learn more about $\mathcal{D}$ by just buying their own data directly than adaptively querying the scoreboard set.

Another feature of our mechanism is that once a round t is over, we can safely release the datasets S_t and T_t used in that round to the public as unrestricted training data. This way, also poor analysts benefit from adaptive queries made by others, as all data is eventually released, and at any given time, a substantial fraction of the data is public. The ratio of public data to validation data can easily be increased 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 hypothesis tests (potentially a large number of them) and identifying statistically significant discoveries while controlling for multiplicity. Because of the complexity of the discovery process, it is often difficult or even impossible 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" [16]). An alternative, easier approach that has gained popularity in recent years, is to require replication of any declared discoveries on new and independent data [3]. Because the new data is used only for replication or validation, it is much easier to control multiplicity and false discovery concerns.

Our everlasting database can be useful in both the discovery and replication phases. To clarify how, we first have to show that the estimation guarantees it provides can be used for multiplicity control in testing. Assume we have a collection of hypothesis tests on functionals of $\mathcal{D}$ with null hypotheses:

$$H_{0i} : \mathbb{E}[q_i] = e_{0i}.$$

We employ our scheme to obtain estimates A_i of $\mathbb{E}[q_i]$. Setting $\alpha = \beta/2$, Theorem (2) guarantees:

$$\sum_i \mathbb{P}_{H_{0i}} \left[\max_i |A_i - e_{0i}| > \tau \right] \leq \alpha,$$

meaning that for any combination of true nulls, the rejection policy *reject if* $|A_i - e_{0i}| > \tau$ makes no false rejections with probability at least $1 - \alpha$, hence we are controlling the family-wise error rate (FWER) at level α.

This is trivially usable in the replication phase, where an entire community (say, type-I diabetes researchers) can share a single replication server while using the everlasting database scheme to guarantee validity. It can also be usable in the discovery phase if the entire analysis can be described through a set of measurements and tests of the form above.

As a concrete example, consider the important task of discovering associations between genes and disease as expressed in genome-wide association studies [8]. This is often expressed as a large number of tests (say 10^6) on association between specific points in the genome and the risk of a disease. With one everlasting database of cases and one of controls, this can be stated as testing H_{0i} that genetic variant i has the same prevalence in cases and controls. So simply estimating these prevalences empirically and rejecting H_{0i} if the difference exceeds 2τ guarantees FWER control at level $\alpha = \beta$ across all past and future use of this database, regardless of adaptivity.

7 Summary

Our primary contribution is in specifying and 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 a manner that guarantees that non-adaptive queries are inexpensive with high probability, and that the potentially high cost of handling adaptivity only falls upon actually adaptive users. Thus we believe that our scheme can form the basis for practical implementations for use in ML competitions and scientific discovery.

Nevertheless, there are several ways that the everlasting database could be improved. Large constants in the cost guarantees are very pessimistic and could likely be reduced with a tighter analysis and a slightly altered pricing scheme. Also, this mechanism only provides answers at a fixed, additive tolerance τ, and only allows statistical queries. We believe it is possible to address these limitations with a more complex mechanism using existing ideas from the adaptive data analysis literature. For example, allowing users to choose a tolerance τ_i or to evaluate other classes of queries.

Our pricing is optimal in terms of the two user types we studied, but it would be interesting to devise mechanisms with other desirable pricing properties. If a user makes a small number of adaptive queries based on many non-adaptive queries, it might be possible to guarantee lower cost using a more sophisticated mechanism. We would also like to avoid the increasing maximum per-query price by charging users for adaptivity a little bit at a time, rather than all at once. Finally, an interesting question is whether it is possible to reveal the price of a query before answering it and charging for it, while still maintaining validity.

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 García-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 Giannoulatou, et al. Genome-wide 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, 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.

[11] 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.

[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.

[13] Andrew Gelman and Eric Loken. The statistical crisis in science. *The American Statistician*, 102(6):460, 2014.

[14] M. Hardt and J. Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS*, pages 454–463, 2014.

[15] Moritz Hardt. Climbing a shaky ladder: Better adaptive risk estimation. *CoRR*, abs/1706.02733, 2017.

[16] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[17] Kobbi Nissim and Uri Stemmer. On the generalization properties of differential privacy. *CoRR*, abs/1504.05800, 2015.

[18] Tom Simonite. Why and how baidu cheated an artificial intelligence test, 2015.

[19] Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *COLT*, pages 1588–1628, 2015.

Relevant Results in Differential Privacy

Here, we state without proof definitions and results from other work which we use in the proof of Lemma 5.

Definition 1. *A randomized algorithm $\mathcal{M} : \mathcal{X}^* \mapsto \mathcal{Y}$ is (ϵ, δ)-differentially private if for all $E \subseteq \mathcal{Y}$ and all datasets $S, S' \in \mathcal{X}^*$ differing in a single element:*

$$\mathbb{P}\left[\mathcal{M}(S) \in E\right] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{M}(S') \in E\right] + \delta.$$

Proposition 1 ([17, 4])**.** *Let $\mathcal{M}$ be an (ϵ, δ)-differentially private algorithm that outputs a function from $\mathcal{X}$ to $[0, 1]$. For a random variable $S \sim \mathcal{D}^n$ we let $q = \mathcal{M}(S)$. Then for $n \geq 2\ln(8/\delta)/\epsilon^2$,*

$$\mathbb{P}\left[\left|\mathcal{E}_S\left[q\right] - \mathbb{E}\left[q\right]\right| \geq 13\epsilon\right] \leq \frac{2\delta}{\epsilon}\ln\left(\frac{2}{\epsilon}\right).$$

Definition 2 (Definition 1.1 [6])**.** *A randomized mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ is ρ-zero-concentrated differentially private (henceforth ρ-zCDP) if, for all $S, S' \in \mathcal{X}^n$ differing on a single entry and all $\alpha \in (1, \infty)$,*

$$D_\alpha\left(\mathcal{M}(S)||\mathcal{M}(S')\right) \leq \rho\alpha,$$

where $D_\alpha\left(\mathcal{M}(S)||\mathcal{M}(S')\right)$ is the α-Rényi divergence between the distribution of $\mathcal{M}(S)$ and $\mathcal{M}(S')$.

Proposition 2 (Proposition 1.6 [6])**.** *Let q be a statistical query. Consider the mechanism $\mathcal{M} : \mathcal{X}^n \to \mathbb{R}$ that on input S, releases a sample from $\mathcal{N}(\mathcal{E}_S\left[q\right], \sigma^2)$. Then $\mathcal{M}$ satisfies $\frac{1}{2n^2\sigma^2}$-zCDP.*

Proposition 3 (Lemma 1.7 [6])**.** *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ and $\mathcal{M}' : \mathcal{X}^n \to \mathcal{Z}$ be randomized algorithms. Suppose $\mathcal{M}$ satisfies ρ-zCDP and $\mathcal{M}'$ satisfies ρ'-zCDP. Define $\mathcal{M}'' : \mathcal{X}^n \to \mathcal{Y} \times \mathcal{Z}$ by $\mathcal{M}''(x) = (\mathcal{M}(x), \mathcal{M}'(x))$. Then $\mathcal{M}''$ satisfies $(\rho + \rho')$-zCDP.*

Proposition 4 (Proposition 1.3 [6])**.** *If $\mathcal{M}$ provides ρ-zCDP, then $\mathcal{M}$ is $\left(\rho + 2\sqrt{\rho\ln(1/\delta)}, \delta\right)$-differentially private for any $\delta > 0$.*