The Limits of Post-Selection Generalization

Kobbi Nissim
Georgetown University
kobbi.nissim@georgetown.edu

Adam Smith†
Boston University
ads22@bu.edu

Thomas Steinke
IBM Research – Almaden
posel@thomas-steinke.net

Uri Stemmer‡
Ben-Gurion University
u@uri.co.il

Jonathan Ullman§
Northeastern University
jullman@ccs.neu.edu

Abstract

While statistics and machine learning offers numerous methods for ensuring generalization, these methods often fail in the presence of post selection—the common practice in which the choice of analysis depends on previous interactions with the same dataset. A recent line of work has introduced powerful, general purpose algorithms that ensure a property called post hoc generalization (Cummings et al., COLT’16), which says that no person when given the output of the algorithm should be able to find any statistic for which the data differs significantly from the population it came from.

In this work we show several limitations on the power of algorithms satisfying post hoc generalization. First, we show a tight lower bound on the error of any algorithm that satisfies post hoc generalization and answers adaptively chosen statistical queries, showing a strong barrier to progress in post selection data analysis. Second, we show that post hoc generalization is not closed under composition, despite many examples of such algorithms exhibiting strong composition properties.

1 Introduction

Consider a dataset $X$ consisting of $n$ independent samples from some unknown population $\mathcal{P}$. How can we ensure that the conclusions drawn from $X$ generalize to the population $\mathcal{P}$? Despite decades of research in statistics and machine learning on methods for ensuring generalization, there is an increased recognition that many scientific findings do not generalize, with some even declaring this to be a “statistical crisis in science” [14]. While there are many reasons a conclusion might fail to generalize, one that is receiving increasing attention is post-selection, in which the choice of method for analyzing the dataset depends on previous interactions with the same dataset. Post-selection can arise from many common practices, such as variable selection, exploratory data analysis, and dataset re-use. Unfortunately, post-selection invalidates traditional methods for ensuring generalization, which assume that the method is independent of the data.

Numerous methods have been devised for statistical inference after post selection (e.g. [16, 18, 12, 13, 23]). These are primarily special purpose procedures that apply to specific types of simple post selection that admit direct analysis. A more limited number of methods apply where the data is reused in one of a small number of prescribed ways (e.g. [2, 4]).

*Supported by NSF award CNS-1565387.
†Supported by NSF awards IIS-1447700 and AF-1763665, a Google Faculty Award and a Sloan Foundation Research Award.
‡Work done while U.S. was a postdoctoral researcher at the Weizmann Institute of Science, supported by a Koshland fellowship, and by the Israel Science Foundation (grants 950/16 and 5219/17).
§Supported by NSF awards CCF-1718088, CCF-1750640, and CNS-1816028, and a Google Faculty Award.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.
A recent line of work initiated by Dwork et al. [9] posed the question: Can we design general-purpose algorithms for ensuring generalization in the presence of post-selection? These works (e.g. [9,8,19,11]) identified properties of an algorithm that ensure generalization under post-selection, including differential privacy [10], information-theoretic measures, and compression. They also identified many powerful general-purpose algorithms satisfying these properties, leading to algorithms for post-selection data analysis with greater statistical power than all previously known approaches.

Each of the aforementioned properties give incomparable generalization guarantees, and allow for qualitatively different types of algorithms. However, Cummings et al. [7] identified that the common thread in each of these approaches is to establish a notion of post hoc generalization (which they originally called robust generalization), and initiated a general study of algorithms satisfying this notion. Informally, an algorithm \( \mathcal{M} \) satisfies post hoc generalization if there is no way, given only the output of \( \mathcal{M}(X) \), to identify any statistical query [17] (that is, a bounded, linear, real-valued statistic on the population) such that the value of that query on the dataset is significantly different from its value on the whole population.

**Definition 1.1** (Post Hoc Generalization [7]). An algorithm \( \mathcal{M} : \mathcal{X}^n \rightarrow \mathcal{Y} \) satisfies \((\varepsilon, \delta)\)-post hoc generalization if for every distribution \( \mathcal{P} \) over \( \mathcal{X} \) and every algorithm \( \mathcal{A} \) that outputs a bounded function \( q : \mathcal{X} \rightarrow [-1, 1] \), if \( X \sim \mathcal{P}^\otimes n \), \( y \sim \mathcal{M}(X) \), and \( q \sim \mathcal{A}(y) \), then \( \mathbb{P} \left[ \left| q(P) - q(X) \right| > \varepsilon \right] \leq \delta \), where we use the notation \( q(P) = \mathbb{E} [q(X)] \) and \( q(X) = \frac{1}{n} \sum_i q(X_i) \), and the probability is over the sampling of \( X \) and any randomness of \( \mathcal{M}, \mathcal{A} \).

Post hoc generalization is easily satisfied whenever \( n \) is large enough to ensure uniform convergence for the class of statistical queries. However, uniform convergence is only satisfied in the unrealistic regime where \( n \) is much larger than \(|\mathcal{X}|\). Algorithms that satisfy post hoc generalization are interesting in the realistic regime where there will exist queries \( q \) for which \( q(P) \) and \( q(X) \) are far, but these queries cannot be found. The definition also extends seamlessly to richer types of statistics than statistical queries. However, restricting to statistical queries only strengthens our negative results.

Since all existing general-purpose algorithms for post-selection data analysis are analyzed via post hoc generalization, it is crucial to understand what we can achieve with algorithms satisfying post hoc generalization. In this work we present several strong limitations on the power of such algorithms. Our results identify natural barriers to progress in this area, and highlight important challenges for future research on post-selection data analysis.

### 1.1 Our Results

**Sample Complexity Bounds for Statistical Queries.** Our first contribution is strong new lower bounds on any algorithm that satisfies post hoc generalization and answers a sequence of adaptively chosen statistical queries—the setting introduced in Dwork et al. [9] and further studied in [11,15,20]. In this model, there is an underlying distribution \( \mathcal{P} \). We would like to design an algorithm \( \mathcal{M} \) that holds a sample \( X \sim \mathcal{P}^\otimes n \), takes statistical queries \( q \), and returns accurate answers \( a \) such that \( a \approx q(P) \). To model post-selection, we consider a data analyst \( \mathcal{A} \) that issues a sequence of queries \( q^1, \ldots, q^k \) where each query \( q^i \) may depend on the answers \( q^1, \ldots, a^{i-1} \) given by the algorithm in response to previous queries.

The simplest algorithm \( \mathcal{M} \) for this task of answering adaptive statistical queries would return the empirical mean \( q^i(X) = \frac{1}{n} \sum \{ q^i(X_i) \} \) in response to each query, and one can show that this algorithm answers each query to within \( \pm \varepsilon \) if \( n \geq \tilde{O}(k/\varepsilon^2) \) samples. Surprisingly, we can improve the sample complexity to \( n \geq \tilde{O}(\sqrt{k}/\varepsilon^2) \) by returning \( q(X) \) perturbed with carefully calibrated noise [9,11]. The analysis of this approach uses post hoc generalization: the noise is chosen so that \( |a - q(X)| \leq \varepsilon/2 \) and the noise ensures \( |q(P) - q(X)| \leq \varepsilon/2 \) for every query the analyst asks.

Our main result shows that the sample complexity \( n = \tilde{O}(\sqrt{k}/\varepsilon^2) \) is essentially optimal for any algorithm that uses the framework of post hoc generalization.

**Theorem 1.2** (Informal). If \( \mathcal{M} \) takes a sample of size \( n \), satisfies \((\varepsilon, \delta)\)-post hoc generalization, and for every distribution \( \mathcal{P} \) over \( \mathcal{X} = \{ \pm 1 \}^{k+O(\log(n/\varepsilon))} \) and every data analyst \( \mathcal{A} \) who asks \( k \) statistical queries, \( \mathbb{P} \left[ \exists j \in [k], \left| q^j(P) - a \right| > \varepsilon \right] \leq \delta \) then \( n = \Omega(\sqrt{k}/\varepsilon^2) \), where the probability is taken over \( X \sim \mathcal{P}^\otimes n \) and the coins of \( \mathcal{M} \) and \( \mathcal{A} \).

To prove our theorem, we construct a joint distribution over pairs \((\mathcal{A}, \mathcal{P})\) such that when \( \mathcal{M} \) is given too small a sample \( X \), and \( \mathcal{A} \) asks \( k - 1 \) statistical queries, then either \( \mathcal{M} \) does not answer all the
queries accurately or $\mathcal{A}$ outputs a $k$-th query $q^*$ such that $q^*(P) - q^*(X) > \varepsilon$. Thus, $\mathcal{M}$ cannot be both accurate and satisfy post hoc generalization.

Our proof of this result refines the techniques in [15][20]—which yield a lower bound of $n = \Omega(\sqrt{k})$ for $\varepsilon = 1/3$.

Our proof circumvents a barrier in previous lower bounds. The previous works use the sequence of queries to uncover almost all of the sample held by the mechanism (a “reconstruction attack” of sorts). Once the analyst has identified all the points in the sample, it is easy to force an error: the analyst randomly asks one of two queries — zero everywhere or zero on the reconstructed sample and one elsewhere — that “look the same to” $\mathcal{M}$ but have different true answers.

We cannot use that approach because in our setting it is impossible to reconstruct any of the sample. Indeed, for the parameter regime we consider, differentially private algorithms could be used to prevent reconstruction with any meaningful confidence. All we can hope for is a weak approximate reconstruction of the sample. This means the algorithm will have sufficient information to distinguish the aforementioned two queries and we cannot end the proof the same way as previously.

Intuitively, our attack approximately reconstructs the dataset in a way that is only $O(\varepsilon)$ better than guessing. This is not enough to completely “cut off” the algorithm and force an error, but, as we will see, does allow the analyst to construct a query $q^*$ that overfits — i.e., $|q^*(X) - q^*(P)| > \varepsilon$.

Our approximate reconstruction is accomplished using a modification of the reconstruction attack techniques of prior work. Specifically, we employ tools from the fingerprinting codes literature [3][22][6] but we output quantitative scores, rather than a hard in/out decision about what is in the sample.

Independently, Wang [24] proved a quantitatively similar bound to Theorem 1.2. However, Wang’s bound only applies to algorithms $\mathcal{M}$ that receive only the empirical mean $q(X)$ of each query (as opposed to the whole data set). This precludes mechanisms such as sample splitting that treat records asymmetrically. Wang’s bound also applies for a slightly different (though closely related) class of statistics.

The dimensionality of $X$ required in our result is at least as large as $k$, which is somewhat necessary. Indeed, if the support of the distribution is $\{\pm 1\}^k$, then there is an algorithm $\mathcal{M}$ that takes a sample of size just $O(\sqrt{d \log(k)/\varepsilon^3})$ [9][11], so the conclusion is simply false if $d \ll k$. Even when $d \ll k$, the aforementioned algorithms require running time at least $2^k$ per query. [15][20] also showed that any polynomial time algorithm that answers $k$ queries to constant error requires $n = \Omega(\sqrt{k})$. We improve this result to have the optimal dependence on $\varepsilon$.

**Theorem 1.3 (Informal).** Assume one-way functions exist and let $c > 0$ be any constant. If $\mathcal{M}$ takes a sample of size $n$, has polynomial running time, satisfies $(\varepsilon, \delta)$-post hoc generalization, and for every distribution $P$ over $X = \{\pm 1\}^k\{1 + O(\log(n/\varepsilon))\}$ and every data analyst $\mathcal{A}$ who asks $k$ statistical queries, $\mathbb{P}[\exists j \in [k], |q^j(P) - a| > \varepsilon] \leq \delta$, then $n = \Omega(\sqrt{k}/\varepsilon^2)$, where the probability is taken over $X \sim P^n$ and the coins of $\mathcal{M}$ and $\mathcal{A}$.

We prove the information-theoretic result (Theorem 1.2) in Section 2. Due to space restrictions, we defer the computational result (Theorem 1.3) to the full version of this work.

**Negative Results for Composition.** Differential privacy provides optimal or near-optimal methods for answering an adaptively-chosen sequence of statistical queries. However, even for answering statistical queries, outside constraints sometimes preclude randomized algorithms (to allay reproducibility concerns, for instance). Furthermore, one of the main goals of the emerging study of adaptive data analysis is to understand unstructured, unplanned dataset re-use.

At this point, we know several techniques for reasoning about generalization in the adaptive setting: differential privacy and algorithmic stability, information bounds, and compression (and there may be many more yet to be discovered) [7]. These techniques are not directly comparable, but they all use posthoc generalization as a fundamental unit of their analysis. If posthoc generalization were to compose well, then this would provide an avenue for combining these techniques (and possibly others). However, we show that this is not the case and, hence, we must search elsewhere for a unifying theory.

Intuitively, we show that, if the same dataset is analyzed by many different algorithms each satisfying post hoc generalization, then the composition of these algorithms may not satisfy post hoc generaliza-
We begin with an information-theoretic lower bound for a class of algorithms \( M \). We are interested in the ability of \( M \) to generate outputs that are natural - that is, not overly complex or noisy. We prove the information-theoretic result (Theorem 1.4) in Section 3. Due to space restrictions, we defer the computational result (Theorem 1.5) to the full version of this work.

Theorem 1.4. For every \( n \in \mathbb{N} \) there is a collection of \( \ell = O(\log n) \) algorithms \( M_1, \ldots, M_\ell \) that take \( n \) samples from a distribution over \( X = \{0, 1\}^{\log n} \) such that (1) each of these algorithms is \( (\epsilon, \delta) \)-post hoc generalizing for every \( \delta > 0 \) and \( \epsilon = O(\sqrt{\log(n/\delta)} / n^{0.009}) \), but (2) the composition \( (M_1, \ldots, M_\ell) \) is not \( (1.999, .999) \)-post hoc generalizing.

If we consider a relaxed notion of computational post hoc generalization, then we show that composition can fail even for just two algorithms. Informally, computational post hoc generalization means that Definition 1.1 is satisfied when the algorithm \( A \) runs in polynomial time.

Theorem 1.5. Assume one-way functions exist. For every \( n \in \mathbb{N} \) there are two algorithms \( M_1, M_2 \) that take \( n \) samples from a distribution over \( X = \{0, 1\}^{\log n} \) such that (1) both algorithms are \( (\epsilon, \delta) \)-computationally post hoc generalizing for every \( \delta > n^{-O(1)} \) and \( \epsilon = O(\sqrt{\log(n/\delta)} / n^{0.009}) \), but (2) the composition \( (M_1, M_2) \) is not \( (1.999, .999) \)-computationally post hoc generalizing.

We defer the computational result (Theorem 1.4) to Section 3. Due to space restrictions, we defer the computational result (Theorem 1.5) to the full version of this work.

2 Lower Bounds for Statistical Queries

2.1 Post Hoc Generalization for Adaptive Statistical Queries

We are interested in the ability of interactive algorithms satisfying post hoc generalization to answer a sequence of statistical queries. Definition 1.1 applies to such algorithms via the following experiment.

Algorithm 1: \( AQ_{X,n,k}[M \rightleftharpoons A] \)

\( A \) chooses a distribution \( \mathcal{P} \) over \( X \)

\( X \sim \mathcal{P}^\otimes n \) and \( X \) is given to \( M \) (but not to \( A \))

For \( j = 1, \ldots, k \)

\( A \) outputs a statistical query \( q^j \) (possibly depending on \( q^1, a^1, \ldots, q^{j-1}, a^{j-1} \))

\( M(X) \) outputs \( a^j \)

Definition 2.1. An algorithm \( M \) is \( (\epsilon, \delta) \)-post hoc generalizing for \( k \) adaptive queries over \( X \) given \( n \) samples if for every adversary \( A \),

\[
\mathbb{P}_{AQ_{X,n,k}[M \rightleftharpoons A]} \left[ \exists j \in [k] \mid q^j(X) - q^j(\mathcal{P}) > \epsilon \right] \leq \delta.
\]

2.2 A Lower Bound for Natural Algorithms

We begin with an information-theoretic lower bound for a class of algorithms \( M \) that we call natural algorithms. These are algorithms that can only evaluate the query on the sample points they are given. That is, an algorithm \( M \) is natural if, when given a sample \( X = (X_1, \ldots, X_n) \) and a statistical query \( q : X' \to [-1, 1] \), the algorithm \( M \) returns an answer \( a \) that is a function only of \( (q(X_1), \ldots, q(X_n)) \). In particular, it cannot evaluate \( q \) on data points of its choice. Many algorithms in the literature have this property. Formally, we define natural algorithms via the game NAQ\(_{X,n,k}[M \rightleftharpoons A] \). This game is identical to AQ\(_{X,n,k}[M \rightleftharpoons A] \) except that when \( A \) outputs \( q^j \), \( M \) does not receive all of \( q^j \), but instead receives only \( q^j_X = (q^j(X_1), \ldots, q^j(X_n)) \).
The proof uses the analyst $A_{\text{NAQ}}$ described in Algorithm 2. For notational convenience, $A_{\text{NAQ}}$ actually asks $k + 1$ queries, but this does not affect the final result.

**Algorithm 2: $A_{\text{NAQ}}$**

**Parameters:** sample size $n$, universe size $N = \frac{8n}{\epsilon}$, number of queries $k$, target accuracy $\epsilon$

Let $P \leftarrow U[|N|]$, $A^1 \leftarrow \emptyset$, and $\tau \leftarrow 9\epsilon \sqrt{2k \log\left(\frac{96n}{\epsilon}\right)} + 1$

For $j \in [k]$
- Sample $p_j \sim U_{[0,1]}$
- For $i \in [N]$
  - Sample $\tilde{q}_i^j \sim \text{Ber}(p_j)$ and let $q_i^j(i) \leftarrow \begin{cases} q_i^j & i \notin A^j \\ 0 & i \in A^j \end{cases}$
- Ask query $q_i^j$ and receive answer $a^j$

For $i \in [N]$
- Let $z_i^j \leftarrow \begin{cases} \text{trunc}_{0.2}(a^j - p^j) \cdot (q_i^j - p^j) & i \notin A^j \\ 0 & i \in A^j \end{cases}$
  - where $\text{trunc}_x(x)$ takes $x \in \mathbb{R}$ and returns the nearest point in $[-3\epsilon, 3\epsilon]$ to $x$.
- Let $A^{j+1} \leftarrow \{ i \in [N] : \sum_{j=1}^k |z_i^j| > 1 - \tau \}$ (N.B. By construction, $A^j \subseteq A^{j+1}$.)

For $i \in [N]$
- Define $z_i \leftarrow \sum_{j=1}^k z_i^j$ and $q_i^* \leftarrow \frac{z_i}{\tau}$
- Let $q^* : [N] \rightarrow [-1, 1]$ be defined by $q^*(i) \leftarrow q_i^*$

In order to prove Theorem 2.2 it suffices to prove that either the answer $a^j$ to one of the initial queries $q^j$ fails to be accurate (in which case $M$ is not accurate, or that the final query $q^*$ gives significantly different answers on $X$ and $P$ (in which case $M$ is not robustly generalizing). Formally, we have the following proposition.

**Proposition 2.3.** For an appropriate choice of $k = \Theta(\epsilon^4 n^2)$ and $n, \frac{1}{\epsilon}$ sufficiently large, for any natural $M$, with probability at least $2/3$, either (1) $\exists j \in [k]$ $|a^j - q^j(P)| > \epsilon$, or (2) $q^*(X) - q^*(P) > \epsilon$, where the probability is taken over the game $\text{NAQ}_{X,n,k}[M \models A_{\text{NAQ}}]$ and $A_{\text{NAQ}}$ is specified by Algorithm 2.

We prove Proposition 2.3 using a series of claims. The first claim states that none of the values $z_i$ are ever too large in absolute value, which follows immediately from the definition of the set $A^j$ and the fact that each term $z_i^j$ is bounded.

**Claim 2.4.** For every $i \in [N]$, $|z_i| \leq \tau$.

The next claim states that, no matter how the mechanism answers, very few of the items not in the sample get “accused” of membership, that is, included in the set $A^j$.

**Claim 2.5 (Few Accusations).** $\Pr(\left| A_k \setminus X \right| \leq \epsilon N/8) \geq 1 - e^{-\Omega(\epsilon n)}$. 

**Proof.** Fix the biases $p^1, ..., p^k$ as well as all the information visible to the mechanism (the query values $\{ q_i^j : i \notin X, j \in [k] \}$, as well as the answers $a^1, ..., a^k$). We prove that the probability of $F$ is high conditioned on any setting of these variables.

The main observation is that, once we condition on the biases $p^j$, the query values at $\{ q_i^j : i \notin X, j \in [k] \}$ are independent with $q_i^j \sim \text{Ber}(p^j)$. This is true because $M$ is a natural algorithm (so it sees
only the query values for points in \( X \) and, more subtly, because the analyst's decisions about how to sample the \( p_j \)'s, and which points in \( X \) to include in the sets \( A_j \), are independent of the query values outside of \( X \). By the principle of deferred decisions, we may thus think of the query values \( \{q^j_i : i \notin X, j \in [k]\} \) as selected after the interaction with the mechanism is complete.

Fix \( i \notin X \). For every \( j \in [k] \) and \( i \notin X \), we have

\[
E \left[ z^j_i \right] = E \left[ \text{trunc}_{\varepsilon}(a^j - p^j) \cdot (q^j_i - p^j) \right] = E \left[ \text{trunc}_{\varepsilon}(a^j - p^j) \right] \cdot E \left[ q^j_i - p^j \right] = 0.
\]

By linearity of expectation, we also have \( E \left[ z_i \right] = E \left[ \sum_{j=1}^{k} z^j_i \right] = 0 \).

Next, note that \( |z^j_i| \leq 3\varepsilon \), since \( \text{trunc}_{\varepsilon}(a^j - p^j) \in [-3\varepsilon, 3\varepsilon] \) and \( q^j_i - p^j \in [0, 1] \). The terms \( z^j_i \) are not independent, since if a partial sum \( \sum_{j=1}^{\ell} z^j_i \) ever exceeds \( \tau \), then subsequent values \( z^j_i \) for \( j > \ell \) will be set to 0. However, we may consider a relaxed sequence given by sums of the terms \( z^j_i = \text{trunc}_{\varepsilon}(a^j - p^j) \cdot (q^j_i - p^j) \) (the difference from \( z^j_i \) is that we use values \( \tilde{q} \) instead of \( q \)).

By Hoeffding’s Inequality, the sum is bounded

\[
O(\sqrt{k \log(1/\varepsilon)}) \text{ with high probability, for every } i \notin X \mathbb{P} \left[ \left| \sum_{j=1}^{k} z^j_i \right| > \varepsilon \sqrt{18k \ln \left( \frac{6k}{\varepsilon} \right)} \right] \leq \frac{\varepsilon}{16}.
\]

By Etemadi’s Inequality, a related bound holds uniformly over all the intermediate sums:

\[
\forall i \notin X \mathbb{P} \left[ \exists \ell \in [k] : \sum_{j=1}^{\ell} z^j_i > 3\varepsilon \sqrt{18k \ln \left( \frac{96}{\varepsilon} \right)} - \tau - 1 \right] \leq 3\mathbb{P} \left[ \sum_{j=1}^{k} z^j_i > \varepsilon \sqrt{18k \ln \left( \frac{96}{\varepsilon} \right)} - \tau - 1 \right] \leq \frac{\varepsilon}{16}.
\]

Finally, notice that by construction, the real scores \( z^j_i \) are all set to 0 when an item is added to \( A_j \), so the sets \( A_j \) are nested (\( A_j \subseteq A_{j+1} \)), and a bound on partial sums of the \( z^j_i \) applies equally well to the partial sums of the \( z^j_i \). Thus, \( \forall i \notin X \mathbb{P} \left[ \exists \ell \in [k] : \sum_{j=1}^{\ell} z^j_i > \tau - 1 \right] \leq \frac{\varepsilon}{16} \)

Now, the scores \( z^j_i \) are independent across players (again, because we have fixed the biases \( p_j \) and the mechanism’s outputs). We can bound the probability that more than \( \frac{nN}{4} \) elements \( i \) are “accused” over the course of the algorithm using Chernoff’s bound: \( \mathbb{P} \left[ |A^k \setminus X| > \frac{\varepsilon}{8} N \right] \leq e^{-\varepsilon N/64} \leq e^{-\Omega(n)} \)

The claim now follows by averaging over all of the choices we fixed.

The next claim states that the sum of the scores over all \( i \) not in the sample is small.

\textbf{Claim 2.6.} \textit{With probability at least } \( \frac{99}{100} \mathbb{P} \sum_{i \in [N] \setminus X} z_i = O(\varepsilon \sqrt{Nk}) \).

\textbf{Proof.} Fix a choice of \( (p_1, \ldots, p_k) \in [0, 1]^k \), the in-sample query values \( (q^1_X, \ldots, q^k_X) \in \{0, 1\}^{n \times k} \), and the answers \( (a^1, \ldots, a^k) \in [0, 1]^k \). Conditioned on these, the values \( z_i \) for \( i \notin X \) are independent and identically distributed. They have expectation 0 (see the proof of Claim 2.5) and are bounded by \( \tau \) (by Claim 2.4). By Hoeffding’s inequality, with probability at least \( \frac{99}{100} \sum_{i \in [N] \setminus X} z_i = O(\tau \sqrt{N}) = O(\varepsilon \sqrt{Nk}) \) as desired. The claim now follows by averaging over all of the choices we fixed.

\textbf{Claim 2.7.} \textit{There exists } \( c > 0 \) \textit{such that, for all sufficiently small } \( \varepsilon \) \textit{and sufficiently large } \( n \), \textit{with probability at least } \( \frac{99}{100} \), \textit{either } \( \exists j \in [k] : |a^j - q^j(P)| > \varepsilon \) \textit{(large error), or } \( \sum_{i \in [N]} z_i \geq ck \) \textit{(high scores in sample)}.
We prove two main statements: first, that these untruncated scores are equal to the truncated ones with high probability as long as the mechanism’s answers are accurate. Second, that the expected sum of untruncated scores is large. This gives us the desired final statement.

To relate the truncated and untruncated scores, consider the following three key events:

1. (“Few accusations”): Let $F$ the event that, at every round $j$, set of “accused” items outside of the sample is small: $|A^j_k \setminus X| \leq \varepsilon N/8$. Since the $A^j$ are nested, event $F$ implies the same condition for all $j$ in $[k]$.

2. (“Low population error”): Let $G$ be the event that at every round $j \in [k]$, the mechanism’s answer satisfies $|a^j - p^j| \leq 3\varepsilon$.

3. (“Representative queries”): Let $H$ be the event that $|\bar{q}^j(P) - p^j| \leq \varepsilon$ for all rounds $j \in [k]$—that is, each query’s population average is close to the corresponding sampling bias $p^j$.

**Sub-Claim 2.9.** Conditioned on $F \cap G \cap H$, the truncated and untruncated scores are equal. Specifically, $|a^j - p^j| \leq 3\varepsilon$ for all $j \in [k]$.

**Proof.** We can bound the difference $|a^j - p^j|$ via the triangle inequality:

$$|a^j - p^j| \leq |a^j - q^j(P)| + |q^j(P) - \bar{q}^j(P)| + |\bar{q}^j(P) - p^j|.$$  

The first term is the mechanism’s sample error (bounded when $G$ occurs). The second is the distortion of the sample mean introduced by setting the query values of $i \in A^j$ to 0. This distortion is at most $|A^j|/N$. When $F$ occurs, $A^j$ has size at most $|X| + |A^j \setminus X| \leq n + \varepsilon N/8 = \varepsilon N/4$, so the second term is at most $\varepsilon/4$. Finally, the last term is bounded by $\varepsilon$ when $H$ occurs, by definition. The three terms add to at most $3\varepsilon$ when $F$, $G$, and $H$ all occur. \[\square\]

We can bound the probability of $H$ via a Chernoff bound: The probability of that a binomial random variable deviates from its mean by $\varepsilon N$ is at most $2 \exp(-\varepsilon^2 N/3)$.

The technical core of the proof is the use of the fingerprinting lemma to analyze the difference $D$ between the sum of untruncated scores and the summed population errors: $D := \sum_{i=1}^N \hat{z}_i - \sum_{j=1}^k |a^j - q^j(P)| - k E \left[ \frac{|A^j|}{N-|A^j|} \right]$.

**Sub-Claim 2.10.** $E[D] = \Omega(k)$

**Proof.** We show that for each round $j$, the expected sum of scores for that round $\sum_i \hat{z}_i^j$ is at least $1/12 - E \left[ |a^j - q^j(P)| - \frac{|A^j|}{N-|A^j|} \right]$. This is true even when we condition on all the random choices and communication in rounds 1 through $j-1$. Adding up these expectations over all rounds gives the desired expectation bound for $D$.  

7
First, note that summing $z_i^j$ over all elements $i \in [N]$ is the same as summing over that round’s unaccused elements $i \in [N] \setminus A^j$ (since $\tilde{z}_i^j = 0$ for $i \in A^j$). Thus,

$$\sum_{i=1}^{N} \tilde{z}_i^j = \sum_{i \in [N] \setminus A^j} \tilde{z}_i^j = (a^j - p^j) \sum_{i \in [N] \setminus A^j} (q_i^j - p_i^j).$$

We can now apply the Fingerprinting Lemma, with $m = N - |A^j|$, $p = p^j$, $x_i = \tilde{q}_i^j$ for $i \notin A^j$, and $f((x_i)_{i \notin A^j}) = a^j$ (note that $f$ depends implicitly on $A_j$, but since we condition on the outcome of previous rounds, we may take $A^j$ as fixed for round $j$). We obtain

$$\mathbb{E} \left[ \sum_{i=1}^{N} \tilde{z}_i^j \right] \geq \frac{1}{12} - \mathbb{E} \left[ a^j - \frac{1}{N - |A^j|} \cdot \sum_{i \notin A^j} q_i^j \right].$$

Now the difference between $\frac{1}{N - |A^j|} \sum_{i \notin A^j} q_i^j$ and the actual population mean $\frac{1}{N} \sum_{i=1}^{N} q_i^j$ is at most $N \cdot (\frac{1}{N} - \frac{1}{N - |A^j|}) = \frac{|A^j|}{|N - |A^j||}$. Thus we can upper-bound the term inside the right-hand side expectation above by $|a^j - q^j(P)| + \frac{|A^j|}{|N - |A^j||}$.

A direct corollary of Sub-Claim 2.10 is that there is a constant $c' > 0$ such that, with probability at least $199/200, D \geq c' k$. Let’s call that event $I$.

Conditioned on $F \cap G \cap H$, we know that each $\tilde{z}_i$ equals the real score $z_i$ (by the first sub-claim above), that $|a^j - q^j(P)| \leq 3\varepsilon$ for each $j$, and that $|A^j| \leq \varepsilon N/8$. If we also consider the intersection with $I$, then we have $D \geq c' k - 3k\varepsilon - k \frac{\varepsilon^2}{8} \geq k(c' - 4\varepsilon)$ (for sufficiently small $\varepsilon$). By a union bound, the probability of $-(F \cap G \cap I)$ is at most $1/200 + \exp(-\Omega(\varepsilon^2 n)) \leq 1/100$ (for sufficiently large $n$). Thus we get $\Pr\left( \neg(G) \lor \left( \sum_{i=1}^{N} z_i \geq \ell k \right) \right) \geq \frac{99}{100}$, where $c = c' - 4\varepsilon$ is positive for sufficiently small $\varepsilon$. This completes the proof of Claim 2.7.

To complete the proof of the proposition, suppose that $|a^j - q^j(P)| \leq \varepsilon$ for every $j$, so that we can assume $\sum_{i \notin X} z_i = \Omega(k)$. Then, we can show that, when $n$ is sufficiently large and $k \geq \varepsilon^4 n^2$, the final query $q^*$ will violate robust generalization. A relatively straightforward calculation (omitted for space) shows that for the query $q^*$ that we defined, $q^*(X) - q^*(P) = \Theta(\varepsilon \sqrt{\ell})$. Now, we choose an appropriate $k = \Theta(\varepsilon^4 n^2)$ we will have that $q^*(X) - q^*(P) > \varepsilon$. By this choice of $k$, the first term in the final line above will be at least $2\varepsilon$. Also, we have $N \geq n = \Theta(\sqrt{\ell} / \varepsilon^2)$, so when $k$ is larger than some absolute constant, the $O(1 / \sqrt{N})$ term in the final line above is $\Theta(\varepsilon / \sqrt{\ell}) \leq \varepsilon$. Thus, by Claims 2.6 and 2.7 either $M$ fails to be accurate, so that $\exists j \in [k] \text{ s.t. } |a^j - q^j(P)| > \varepsilon$, or we find a query $q^*$ such that $q^*(X) - q^*(P) > \varepsilon$.

### 2.3 Lower Bounds for All Algorithms via Random Masks

We prove Theorem 1.2 by constructing the following transformation from an adversary that defeats all natural algorithms to an adversary that defeats all algorithms. The main idea of the reduction is to use random masks to hide information about the evaluation of the queries at points outside of the dataset, which effectively forces the algorithm to behave like a natural algorithm because, intuitively, it does not know where to evaluate the query apart from on the dataset. The reduction is described in Algorithm 3. Due to space restrictions, we omit its analysis due to space.

### 3 Post Hoc Generalization Does Not Compose

In this section we prove that post hoc generalization is not closed under composition.

**Theorem 3.1.** For every $n \in \mathbb{N}$ and every $\alpha > 0$ there is a collection of $\ell = O\left(\frac{1}{\alpha} \log n\right)$ algorithms $M_1, \ldots, M_\ell : \{\{0, 1\}^5 \log n\}^n \to Y$ such that (1) for every $i = 1, \ldots, \ell$ and $\delta > 0$, $M_i$ satisfies $(\varepsilon, \delta)$-post hoc generalization for $\varepsilon = O\left(\sqrt{\log(n/\delta) / n^{1-\alpha}}\right)$, but (2) the composition $(M_1, \ldots, M_\ell)$ is not $(2 - \frac{2}{n^{2/3}}, 1 - \frac{1}{n^{\alpha/2}})$-post hoc generalizing.

The result is based on an algorithm that we call Encrypermute. Before proving Theorem 3.1, we introduce Encrypermute and establish the main property that it satisfies. The key facts about Encrypermute are as follows.
Algorithm 3: $\mathcal{A}_{\text{NAQ}}$

Parameters: sample size $n$, universe size $N = \frac{8n}{\varepsilon}$, number of queries $k$, target accuracy $\varepsilon$.
Oracle: an adversary $\mathcal{A}_{\text{NAQ}}$ for natural algorithms with sample size $n$, universe size $N$, number of queries $k$, target accuracy $\varepsilon$.

Let $X = \{(i, y)\}_{1 \leq i \leq N, y \in \{\pm 1\}^k}$

For $i \in [N]$
- Choose $m_i = (m_i^1, \ldots, m_i^k) \sim U(\{\pm 1\}^k)$

Let $\mathcal{P}$ be the uniform distribution over pairs $(i, m_i)$ for $i \in [N]$

For $j \in [k]$
- Receive the query $\hat{q}^j : [N] \to \{\pm 1\}$ from $\mathcal{A}_{\text{NAQ}}$
- Form the query $q^j(i, y) = y^j \oplus m_i^j \oplus \hat{q}^j(i)$ (NB: $q^j(i, m_i) = \hat{q}^j(i)$)
- Send the query $q^j$ to $\mathcal{M}$ and receive the answer $a^j$
- Send the answer $a^j$ to $\mathcal{A}_{\text{NAQ}}$

Algorithm 4: Encrypermute

Input: Parameter $k$, and a sample $X = (x_1, x_2, \ldots, x_n) \in (\{0, 1\}^d)^n$ for $d = 5 \log n$.

If $X$ contains $n$ distinct elements
- Let $\pi$ be the permutation that sorts $(x_1, \ldots, x_k)$ and identify $\pi$ with $r \in \{0, 1, \ldots, k! - 1\}$
- Let $\alpha \in [0, 1]$ be the largest number such that $k \geq n^\alpha$ and let $t \leftarrow \alpha k/20$ (NB: $2^{4t} \leq k!$)
- Identify $(x_{k+1}, \ldots, x_{k+t}) \in (\{0, 1\}^d)^t$ with a number $m \in \{0, 1, \ldots, k! - 1\}$
- Return $c = m + r \mod k!$

Else
- Return a random number $c \in \{0, 1, \ldots, k! - 1\}$

Claim 3.2. Let $\mathcal{D}$ be any distribution over $(\{0, 1\}^d)^n$. Let $D \sim \mathcal{D}$, let $X$ be a random permutation of $D$, and let $C \leftarrow \text{Encrypermute}(X)$. Then $D$ and $C$ are independent.

Intuitively, the claim follows from the fact that $r$ is uniformly random and depends only on the permutation, so it is independent of $D$. Therefore $m + r \mod k!$ is random and independent of $m$.

Lemma 3.3. $\forall \delta > 0$, Encrypermute satisfies $(\varepsilon, \delta)$-post hoc generalization for $\varepsilon = \sqrt{2 \log(2/\delta)}/n$.

Intuitively the lemma follows from the fact that $C$ is independent of $D$. We omit the proof of both of these claims due to space restrictions.

Proof of Theorem [3.7] Fix $\alpha \in (0, 1)$, and let $\mathcal{M}_1$ denote the mechanism that takes a database of size $n$ and outputs the first $n^\alpha$ elements of its sample. As $\mathcal{M}_1$ outputs a sublinear portion of its input, it satisfies post hoc generalization with strong parameters. Specifically, by [2] Lemma 3.5, $\mathcal{M}_1$ is $(\varepsilon, \delta)$-post hoc generalizing for $\varepsilon = O\left(\sqrt{\log(n/\delta)/n^{1-\alpha}}\right)$.

Now consider composing $\mathcal{M}_1$ with $O\left(\frac{1}{\alpha} \log n\right)$ copies of Encrypermute, with exponentially growing choices for the parameter $k$, where for the $i$th copy we set $k = (1 + \frac{n}{2m})^i \cdot n^\alpha$. By Lemma 3.3, each of these mechanisms satisfies post hoc generalization for $\varepsilon = O(\sqrt{\log(1/\delta)/n})$, so this composition satisfies the assumptions of the theorem.

Let $\mathcal{P}$ be the uniform distribution over $\{0, 1\}^d$, where $d = 5 \log n$, and let $X \sim \mathcal{P}^\otimes n$. By a standard analysis, $X$ contains $n$ distinct elements with probability at least $\left(1 - \frac{1}{2n}\right)^n$. Assuming that this is the case, we have that the first copy of Encrypermute outputs $c = m + r \mod k!$, where $m$ encodes the rows of $X$ in positions $n^\alpha + 1, \ldots, (1 + \frac{n}{2m})n^\alpha$, and where $r$ is a deterministic function of the first $n^\alpha$ rows of $X$. Hence, when composed with $\mathcal{M}_1$, these two mechanism reveal the first $(1 + \frac{n}{2m})n^\alpha$ rows of $X$. By induction, the output of the composition of all the copies of Encrypermute with $\mathcal{M}_1$ reveals all of $X$. Hence, from the output this composition, we can define the predicate $q : \{0, 1\}^d \to \{\pm 1\}$ that evaluates to 1 on every element of $X$, and to -1 otherwise. This predicate satisfies $q(X) = 1$ but $q(\mathcal{P}) \leq -1 + 2n/2^d = -1 + 2/n^4$. \hfill \square
References


