# Controlling Bias in Adaptive Data Analysis Using Information Theory

**Daniel Russo**
Microsoft Research

**James Zou**
Microsoft Research

## Abstract

Modern big data settings often involve messy, high-dimensional data, where it is not clear *a priori* what are the right questions to ask. To extract the most insights from a dataset, the analyst typically needs to engage in an iterative process of adaptive data analysis. The choice of analytics to be performed next depends on the results of the previous analyses on the same data. It is commonly recognized that such adaptivity (also called researcher degrees of freedom), even if well-intentioned, can lead to false discoveries, contributing to the crisis of reproducibility in science. In this paper, we propose a general information-theoretic framework to quantify and provably bound the bias of *arbitrary* adaptive analysis process. We prove that our mutual information based bound is tight in natural models. We show how this framework can give rigorous insights into when commonly used feature selection protocols (e.g. rank selection) do and do not lead to biased estimation. We also show how recent insights from differential privacy emerge from this framework when the analyst is assumed to be adversarial, though our bounds applies in more general settings. We illustrate our results with simple simulations.

## 1 Introduction

Modern big data is messy and high-dimensional, and it is often not clear a priori what is the right analysis to perform on the data. To extract the most insight, the analyst typically needs to perform exploratory analysis to make sense of the data and at the same time test various hypotheses. This is invariably an adaptive process: patterns in the data observed in the first stages of analyses inform what tests are run next and the process iterates. This is also called "researcher degrees of freedom" [17]. Such "data-dredging" is largely maligned by classical statistical theory. It is known that this process, even when the analyst is well-intentioned, can lead to false discovery or large bias in the reported estimates.

Although adaptivity is ubiquitous in data science, it is largely outside the realm of classical statistics. Standard tools of multiple hypothesis testing and false discovery rate (FDR) assume that all the hypotheses to be tested are chosen independently of the dataset. But while any adaptivity renders classical statistical theory invalid, folklore and experience also suggest that not all types of adaptive analysis are equally at risk for false discoveries. In this work, we undertake a general and systematic study into the degree of bias introduced by different forms of adaptivity, in which the choice of which function of the data to report is made *after* observing and analyzing the dataset. Our main result is an information theoretic bound on the bias of an arbitrary, adaptively chosen function of the data. This bound provides a quantitative measure of researcher degrees of freedom, and offers a single lens through which we investigate different forms of adaptivity.

### 1.1 Preview of the main result

We consider a general framework in which a dataset $D$ is drawn from a probability distribution $\mathcal{P}$ over $\mathcal{D}$. Let $\phi_1, ..., \phi_m : \mathcal{D} \to \mathbb{R}$ denote the set of all analyses that the analyst may want to run on the data. The number, $m$, of the $\phi_i$'s is finite but it could be arbitrarily large; in particular $m$ can be exponential in the number of samples in the dataset. Each $\phi_i$ is a random variable that depends on the particular realization of the data $D \sim \mathcal{P}$. After observing $D$ or some summary statistics of $D$, the analyst chooses

Appearing in Proceedings of the $19^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41. 

to report the value $\phi_T(D)$ for $T \in \{1, ..., m\}$. For simplicity, we focus on the case where exactly one $\phi_T$ is selected and reported, but our results can be extended to when multiple $\phi_i$'s are selected. The selection rule $T : \mathcal{D} \to \{1, ..., m\}$ captures how the analyst uses the data and chooses which result to report. Because the choice $T$ is itself a function of the realization $D$, the reported value $\phi_T(D)$ may be significantly biased. For example, $\mathbf{E}[\phi_T(D)]$ could be very far from zero even if each fixed function $\phi_i(D)$ has zero mean.

**Example 1.** A standard analytics setting is when $D \in \mathbb{R}^{n \times d}$; there are $n$ samples and each sample is a $d$-dimensional feature vector. For example, a sample could be an individual and the $d$ features are the genotype of the individual at $d$ loci in his/her genome. Given a vector of phenotypes associated with each sample (e.g. the blood pressure of each individual), a common analysis is to regress each genetic locus against the phenotype. Here $\phi_i$ is the estimated regression coefficient or effect size of the $i$th locus on the phenotype and $m = d$. In most studies, $d$ is in the range of $10^5 - 10^6$ and $n$ is $10^3$. $T$ is the the analyst's selection rule for deciding which set of loci to report as relevant for the phenotype. One common selection rule $T$ is to rank the $\phi_i$'s by absolute value and set $T$ to be the indices corresponding to the top $K$ largest $\phi_i$'s, where $K$ is a fixed constant set ahead of time. Another selection rule is to set a $p$-value threshold and select all the $\phi_i$'s whose $p$-value is smaller than the threshold. Both selection rules suffer from *Winner's Curse* [13], which is the empirical phenomenon whereby the estimated effect size $\phi_i$, where $i$ is selected (i.e. $T = i$), is larger than the true effect size of that locus. This selection bias occurs because the selection policy relies heavily on the realized values $\phi_i$. There are many other selection policies. For example, the analyst may run Lasso on all the loci and $T$ correspond to the subset of loci selected by Lasso. We would like to have a unified framework to understand and bound the selection bias of any policy $T$.

**Example 2.** In the previous example, the selection procedure explicitly uses the values of the $\phi_i$'s. In many adaptive data analysis, the selection procedure uses other information contained in $D$. As before, let $D \in \mathbb{R}^{n \times d}$. Another common goal is to find a pair of features that are highly correlated with each other. Here the $\phi_i$'s are all the pairwise correlation values and $m = d(d-1)/2$. The analyst may perform clustering or principal component analysis (PCA) on $D$ and based on the output of these exploratory analyses select which correlation pair $\phi_i$ to measure on the data and report. This selection procedure $T$ uses the data $D$ but not the values of $\phi_i$'s explicitly, and intuitively it could result in smaller selection bias. The analyst may also measure just one $\phi_{i_1}$, and based on its value, select another pair to measure, $\phi_{i_2}$, and so on. In the end, she has explicitly measured only a few $\phi_i$'s but they were chosen adaptively. We would also like a unified framework to understand the selection bias of these procedures.

In this paper, we bound the degree of bias in terms of an information–theoretic quantity: the mutual information between the choice $T$ of the statistic to report, and the actual realized value of the statistics $(\phi_1(D), ..., \phi_m(D))$. We state this result in a general framework, where $\boldsymbol{\phi} = (\phi_1, ..., \phi_m) : \Omega \to \mathbb{R}^m$ and $T : \Omega \to \{1, .., m\}$ are any random variables defined on a common probability space. Let $\boldsymbol{\mu} = (\mu_1, ..., \mu_m) \triangleq \mathbf{E}[\boldsymbol{\phi}]$ denote the mean of $\boldsymbol{\phi}$. Recall that a real-valued random variable $X$ is $\sigma$–sub-Gaussian if for all $\lambda \in \mathbb{R}$, $\mathbf{E}[e^{\lambda X}] \leq e^{\lambda^2 \sigma^2 / 2}$ so that the moment generating function of $X$ is dominated by that of a normal random variable. Zero–mean Gaussian random variables are sub-Gaussian, as are bounded random variables.

**Proposition 1.** *Suppose that for each $i \in \{1, ..., m\}$, $\phi_i - \mu_i$ is $\sigma$–sub-Gaussian. Then,*

$$|\mathbf{E}[\phi_T] - \mathbf{E}[\mu_T]| \leq \sigma \sqrt{2 I(T; \boldsymbol{\phi})}$$

*where $I$ denotes mutual information.*

The expectation $\mathbf{E}[\phi_T]$ is taken jointly over the randomness in the realized values of $\phi_1, .., \phi_m$ and the selection procedure $T$. It is the expected value of the reported test statistic when selection and estimation are performed on the *same* dataset. The expectation $\mathbf{E}[\mu_T]$ is taken only over the selection procedure $T$, since $\mu$ is a property of the distribution and not of the realized data $D$. This is the expected value of the reported test statistic when, after selection, a *fresh* dataset is used for estimation. The difference $|\mathbf{E}[\phi_T] - \mathbf{E}[\mu_T]|$ quantifies the bias due to the analyst's selection process $T$.

The randomness of $\boldsymbol{\phi}$ is due to the randomness in the realization of the data $D \sim \mathcal{P}$. This captures how each test statistics $\phi_i$ varies if a replication dataset is collected, and hence captures the *noise* in the statistics. The mutual information $I(T; \boldsymbol{\phi})$ then quantifies the dependence of the selection process on the noise in the test statics. Intuitively, a selection process that is more sensitive to the noise (high $I$) is at a greater risk for bias.

Proposition 1 gently interpolates between two extreme cases. If $T$ is independent of $\boldsymbol{\phi}$, $I(T; \boldsymbol{\phi}) = 0$

and therefore $\mathbf{E}\phi_T = \mathbf{E}\mu_T$. It makes sense that there is no bias because the selection does not depend on the actual values of the statistics. If $T = \arg\max_{1\leq i\leq m}\phi_i$, then $I(T;\boldsymbol{\phi})$ is the entropy of $T$, $H(T) = \log(m)$, and we have that $\mathbf{E}[\phi_T - \mu_T] \leq \sigma\sqrt{2\log(m)}$. This is the well known inequality for the maximum of sub-Gaussian random variables. The coming sections will explore cases in between these two extremes in which this bound has interesting implications. Note that while Prop. 1 focuses on bias, Section 2 provides mutual information based guarantees for other measures of the accuracy of the reported statistic $\phi_T$.

We will frequently apply our results when $\phi_i = n^{-1}\sum_{j=1}^n f_i(X_j)$ is the sample average of some function $f_i$ based on an iid sequence $X_1, ..., X_n$. Note that if $f_i(X_j) - \mathbf{E}[f_i(X_j)]$ is sub-Gaussian with parameter $\sigma$, then $\phi_i - \mu_i$ is sub-Gaussian with parameter $\sigma/\sqrt{n}$ and therefore

$$|\mathbf{E}[\phi_T] - \mathbf{E}[\mu_T]| \leq \sigma\sqrt{\frac{2I(T;\boldsymbol{\phi})}{n}}.$$

### 1.2 Related work

Our paper relates to a large body of work on methods for providing meaningful statistical inference and preventing false discovery. Much of this literature has focused on controlling the false discovery rate [1, 2] in multiple-hypothesis testing where the hypotheses are not adaptively chosen. Another line of work studies confidence intervals and significance tests for parameter estimates in sparse high dimensional linear regression (see [12, 15, 20] and the references therein). In learning theory, PAC-Bayes analysis gives powerful generalization bounds in terms of KL-divergence [16]. These techniques bear a resemblance to the results in this paper, and it would be interesting to explore connections between PAC-Bayes theory and adaptive data analysis.

One recent line of work [9, 18] proposes a framework for assigning significance and confidence intervals in selective inference, where model selection and significance testing are performed on the same dataset. These papers propose controlling the probability of error conditioned on the event that the model was chosen. While some extremely powerful results can be derived in the selective inference framework (e.g. [19]), it requires that the conditional distribution $\mathbf{P}(\phi_i = \cdot|T = i)$ is known and can be directly analyzed. This requires that the candidate models and the selection procedure $T$ are mathematically tractable and specified by the analyst beforehand. Our approach gives up some of the sharpness of the selective inference results, but it enables us to formalize insights that applies to general analysis procedures.

Another recent line of work in computer science [3, 7, 8, 11] has established a powerful connection between adaptive data analysis and differential privacy. In their framework, the analyst interacts with a dataset indirectly, and sees only the noisy output of a differentially private mechanism. In Section 4, we specialize our results to this setting, and show that such a mechanism controls the mutual information $I(T;\boldsymbol{\phi})$. The results from this literature are designed for worst-case, adversarial data analysts. We provide guarantees that vary with the selection rule, but apply to all possible selection procedures, including ones that are non-differentially private.

**Outline.** In Sec. 2, we flesh out the main result, Proposition 1, and show that it is tight by proving a matching lower bound on bias. We extend it to sub-exponential random variables and p values, and show that it directly gives new estimates of order statistics. Sec. 3 applies our framework to analyze several basic selection protocols and illustrates when they do and do not lead to bias. Sec. 4 considers a general model of an adaptive data analyst and discusses the information budget framework. It also discusses connections to differential privacy.

## 2 Unpacking the main result

This section establishes additional general results linking the price of adaptivity in data analysis to the mutual information $I(T;\boldsymbol{\phi})$. We first present a lower bound showing that procedures that use a lot of information can have bias as large as $\sigma\sqrt{2I(T;\boldsymbol{\phi})}$. We then provide an analogue of Proposition 1 that applies to sub-exponential random variables, and show that mutual information controls other measures of the cost of adaptivity, such as the average squared prediction error $\mathbf{E}[(\phi_T - \mu_T)^2]$. The final subsection treats situations in which $\phi_1, .., \phi_m$ are p-values corresponding to $m$ different hypothesis tests.

### 2.1 Matching Lower Bound

This subsection provides a lower bound that matches Proposition 1, including constants. This shows that Proposition 1 is tight.

**Proposition 2.** *Let $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3, ...)$ be a collection of independent normally distributed random variables with mean 0 and variance $\sigma^2$. For $B > 1$, let $T_B = \arg\max_{1\leq i\leq \lfloor e^B\rfloor} \phi_i$. Then $I(T_B;\boldsymbol{\phi}) \leq B$ and*

$$\mathbf{E}[\phi_{T_B}] - \sigma\sqrt{2B} \longrightarrow 0$$

*as $B \to \infty$. In addition, there is a universal numerical constant $c > 0$ such that*

$$\mathbf{E}[\phi_{T_B}] \geq c\sigma\sqrt{2B} \qquad \forall B \geq 2.$$

### 2.2 Order Statistics of Gaussians

The form of $I(T;\boldsymbol{\phi})$ makes it particularly easy to analyze selection policies that depend on the order statistics of $\phi_i$. We have seen that in the case of the maximal statistic the bound is tight. As another illustration, consider the policy that first orders $\phi_i$ by size and then uniformly randomly selects from the largest $m_0$ $\phi_i$'s. Let $T$ denote the index of the randomly chosen element, and let $\phi_{(1)} > \phi_{(2)} > ... > \phi_{(m)}$ denote the values of $\phi_i$ sorted from the largest to the smallest. We immediately have $\mathbf{E}\left[\frac{1}{m_0}\sum_{i=1}^{m_0}\phi_{(i)}\right] = \mathbf{E}[\phi_T] \leq \sigma\sqrt{2(H(T) - H(T|\boldsymbol{\phi}))} = \sigma\sqrt{2(\log m - \log m_0)} = \sigma\sqrt{2\log\frac{m}{m_0}}$. For example, if $m_0 = m^\alpha$ for $\alpha < 1$, then this implies $\mathbf{E}[\phi_T] \leq \sigma\sqrt{2(1-\alpha)\log(m)}$. We have not seen this formula for $\mathbf{E}\left[\frac{1}{m_0}\sum_{i=1}^{m_0}\phi_{(i)}\right]$ in literature, though it would not be surprising if it is known under other contexts.

This bound is also tight as $m/m_0 \to \infty$. For convenience, we show this when $m$ is divisible by $m_0$. Consider the following alternative selection policy $\hat{T}$. Randomly partition the $\phi_i$'s into $m_0$ groups of size $m/m_0$. Within each group, select the maximal $\phi_i$ and from these $m_0$ maximal $\phi_i$'s randomly select one as $\phi_{\hat{T}}$. Because the average among the $m_0$ group leaders is less than the average among the $\phi_{(1)}, ..., \phi_{(m_0)}$, we have $\mathbf{E}[\phi_{\hat{T}}] \leq \mathbf{E}[\phi_T]$. Moreover, each group leader converges to $\sigma\sqrt{2\log m/m_0}$ and since the groups are independent, the average also converges to $\sigma\sqrt{2\log m/m_0}$.

### 2.3 Subexponential random variables

Although sub-Gaussian distributions are commonly used, it is useful to have results that apply to random variables with somewhat heavier tails. Here we derive an analogue of Proposition 1 that applies to the broader class of sub-exponential distributions.

**Definition 1.** A random variable $X$ with mean $\mu = \mathbf{E}[X]$ is sub-exponential if there are non-negative parameters $(\sigma, b)$ such that

$$\mathbf{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2} \text{ for all } |\lambda| < \frac{1}{b}.$$

**Proposition 3.** *Suppose that for each $i \in \{1, ..., m\}$, $\phi_i - \mu_i$ is sub-exponential with parameters $(\sigma, b)$. Then*

$$\mathbf{E}[\phi_T - \mu_T] \leq bI(T;\boldsymbol{\phi}) + \frac{\sigma^2}{2b}.$$

*Moreover, if $b \geq 1$, we also have*

$$\mathbf{E}[\phi_T - \mu_T] \leq \sqrt{b}I(T;\boldsymbol{\phi}) + \frac{\sigma^2}{2\sqrt{b}}.$$

This bound on sub-exponential random variables is also tight. Let $\phi_i$ be independent chi-squared distributions, $\chi_1^2$, and let $T$ be the policy for selecting the $\phi_i$ with the maximal value. Since $\chi_1^2$ is sub-exponential with $b = 2$, our bound on $\mathbf{E}[\max_i \phi_i]$ is $2\log n + o(1)$. On the other hand, extreme value theory tells us that $a_n(\max_i \phi_i - b_n) \to \Gamma$, where $a_n = 1/2$, $b_n = 2\log n + o(\log n)$ and $\mathbf{P}(\Gamma \leq x) = e^{-e^{-x}}$. We see that our bound of $2\log n$ is actually tight.

### 2.4 Beyond Bias

While we often focus on the bias $\mathbf{E}[\phi_T] - \mathbf{E}[\mu_T]$, here we note that our techniques allow us to control other properties of $(\phi_T - \mu_T)$. For example, the following corollary of Proposition 1 controls the mean squared distance between $\phi_T$ and $\mu_T$.

**Corollary 1.** *Suppose that for each $i \in \{1, .., m\}$, $-C \leq \phi_i - \mu_i \leq C$ almost surely. Then,*

$$\mathbf{E}[(\phi_T - \mu_T)^2 - V_T] \leq C\sqrt{2I(T;\boldsymbol{\phi})}$$

*where $V_i \triangleq \mathbf{E}[(\phi_i - \mu_i)^2]$ is the variance of $\phi_i$.*

Let $D(\mathbb{P}||\mathbb{Q})$ denote the KL-divergence between probability distributions $\mathbb{P}$ and $\mathbb{Q}$. This result will be used in the next subsection when analyzing the probability of reporting small p-values. Here $\tilde{\phi}_T$ can be thought of as the selected test statistic evaluated on a fresh dataset.

**Proposition 4.** *Let $\tilde{\boldsymbol{\phi}}$ denote a random variable drawn from the marginal distribution of $\boldsymbol{\phi}$, but drawn independently of $T$ and $\boldsymbol{\phi}$. Then*

$$D\left(\mathbf{P}(\phi_T = \cdot)\,||\,\mathbf{P}(\tilde{\phi}_T = \cdot)\right) \leq I\left(T;\boldsymbol{\phi}\right)$$

### 2.5 The Probability of Small p–values

Consider choosing to report the p-value $\phi_T$ corresponding to a single hypothesis test from among a large collection of $\phi_1, ..., \phi_m$ of observed p-values. Under the null hypothesis, each p-value $\phi_i$ is uniformly distributed, so $\mathbf{P}(\phi_i \leq \epsilon) = \epsilon$ for each $\epsilon \in [0, 1]$. Suppose the data analyst rejects the null hypothesis corresponding to $T$ whenever $\phi_T \leq .05$. If $T$ is chosen adaptively so that $\phi_T$ is the smallest p-value among $\phi_1, ..\phi_5$, then the probability of falsely

rejecting the null hypothesis is $1-(.95)^5 \approx .23$. Therefore, at a significance level of .05, even fairly mild forms of adaptivity can create a substantial risk of false discovery. Nevertheless, we argue in this section that very small p-values are very unlikely unless the mutual information $I(T;\boldsymbol{\phi})$ is large.

To build intuition, imagine that $\phi_1, ..., \phi_m \overset{iid}{\sim} \text{Uniform}(0,1)$. If the hypothesis $T = \arg\min_{i \leq m} \phi_i$ with the smallest p-value is selected, the reported p-value is expected to be of order $1/m$. In particular, $\mathbf{E}[\phi_T] = 1/(m+1)$, and

$$\mathbf{P}\left(\phi_T \leq \frac{1}{m}\right) = 1 - \left(1 - \frac{1}{m}\right)^m \longrightarrow 1 - \frac{1}{e}.$$

Therefore, when selecting among $m \approx e^B$ hypotheses, one expects to observe p-values as small as $\epsilon \approx e^{-B}$ but not smaller. Our next proposition extends this line of reasoning, and replaces $B = \log(m)$ with the mutual information between $T$ and $\boldsymbol{\phi}$. It shows that when $\phi_1, ..., \phi_m$ are uniformly distributed, but not necessarily independent, one is very unlikely to observe a p value $\phi_T$ much smaller than $e^{-I(T;\boldsymbol{\phi})}$. under an arbitrary adaptive selection procedure $T$.

**Proposition 5.** *Define* $Z_{\epsilon,i} = \mathbf{1}(\phi_i < \epsilon)$ *and let* $\mathbf{Z}_\epsilon = (Z_{\epsilon,1}, ..., Z_{\epsilon,m})$. *If* $\phi_i \sim \text{Uniform}(0,1)$ *for all* $i \in \{1, .., m\}$ *then*

$$\mathbb{P}(p_T < \epsilon) \leq \epsilon + \sqrt{\frac{I(T;\mathbf{Z}_\epsilon)}{\log(1/2\epsilon)}} \leq \epsilon + \sqrt{\frac{I(T;\phi)}{\log(1/2\epsilon)}}.$$

### 2.6 Regret analysis and value of information

Consider a general problem of optimization under uncertainty. A decision-maker would like to choose the action $x$ from a finite set $\mathcal{X}$ that solves $\max_{x \in \mathcal{X}} f_\theta(x)$. Here $\theta$ is an unknown parameter that is drawn from a prior distribution over a set of possible parameters $\Theta$. We consider the decision-maker's expected shortfall in performance due to not knowing the parameter $\theta$:

$$\mathbf{E}[\max_{x \in \mathcal{X}} f_\theta(x)] - \max_{x \in \mathcal{X}} \mathbf{E}[f_\theta(x)].$$

This measures the value of perfect information about $\theta$: the expected improvement in decision quality that would result from resolving uncertainty about the identity of $\theta$. This is sometimes called the *Bayes risk* or *Bayesian regret* of the decision $\arg\max_{x \in \mathcal{X}} \mathbf{E}[f_\theta(x)]$.

Our main result provides an information theoretic bound on Bayes risk. Let $X^* \in \arg\max_{x \in \mathcal{X}} f_\theta(x)$ denote a true maximizer of the function $f_\theta$. Here $X^*$ is a random variable, since $\theta$ is random, and $X^*$ is a function of $\theta$. Let $\mu(x) = \mathbf{E}[f_\theta(x)]$.

**Proposition 6.** *If for each for each* $x \in \mathcal{X}$, $f_\theta(x) - \mu(x)$ *is* $\sigma$ *sub-Gaussian, then*

$$\mathbf{E}[\max_{x \in \mathcal{X}} f_\theta(x)] - \max_{x \in \mathcal{X}} \mu(x) \leq \sigma\sqrt{2H(X^*)}$$

## 3 Selective inference

In this section, we consider several simple but commonly used procedures of feature selection and parameter estimation. In many applications, such feature selection and estimation are performed on the same dataset. Our information theoretic bound provides a unified framework to understand selection bias in these settings. Our results inform when these these procedure do lead to selection bias and when they do not introduce bias. The key idea is to understand which structures in the data and the selection procedure make the mutual information $I(T;\boldsymbol{\phi})$ significantly smaller than the worst case $\log(m)$. We give several simulation experiments as illustration.

### 3.1 Variance based selection

Imagine that $T$ is chosen after observing some dataset $D$. This dataset determines the values of $\phi_1, ..., \phi_m$, but may also contain a great deal of other information. Manipulating the mutual information shows $I(T;\boldsymbol{\phi}) = H(T) - H(T|\boldsymbol{\phi}) \leq H(T) - I(T;D|\boldsymbol{\phi}) = (1-\alpha)H(T)$ where $\alpha = I(T;D|\boldsymbol{\phi})/H(T)$ captures the fraction of the uncertainty in $T$ that is explained by the data in $D$ beyond the values $\phi_1, ..., \phi_m$. In many cases, instead of being a function of $\boldsymbol{\phi}$, the choice $T$ is a function of data that is more loosely coupled with $\boldsymbol{\phi}$, and therefore we expect that $I(T;\boldsymbol{\phi})$ is much smaller than $H(T)$ (which itself can be less than $\log(m)$).

One setting when the selection of $T$ depends on the statistics of $D$ that are only loosely coupled with $\boldsymbol{\phi}$ is variance based feature selection [14, 21]. Suppose we have $n$ samples and $m$ bio-markers. Let $X_{i,j}$ denote the value of the $i$-th bio-marker on sample $j$. Here $D = \{X_{i,j}\}$. Let $\phi_i = n^{-1}\sum_{j=1}^n X_{i,j}$ be the empirical mean values of the $i$-th biomarker. We are interested in identifying the markers that show significant non-zero mean. A filtering step in many studies is to select markers that have high variance and remove the rest. The rational is that markers that do not vary could be measurement errors or are likely to be less important. A natural question is whether such variance filtering introduces bias.

In our framework, variance selection is exemplified by the selection rule $T = \underset{1 \leq i \leq m}{\arg\max} V_i$ where $V_i = \sum_{j=1}^n (X_{i,j} - \phi_i)^2$. Here we consider the extreme case where only the marker with the largest variance is

selected; all the discussion applies to softer selection when we select the top $K$ markers with the largest variance. The resulting bias is $\mathbf{E}[\phi_T - \mu_T]$. Proposition 1 states that variance selection has low bias if $I(T;\boldsymbol{\phi})$ is small, which is the case if the empirical means and variances, $\phi_i$ and $V_i$, are not too dependent. Indeed, when $X_{i,j}$ are i.i.d. Gaussian samples, $\phi_1, ..., \phi_m$ are independent of $V_1, ..., V_m$ . Therefore $I(T;\boldsymbol{\phi}) = 0$ and we can guarantee that there is no bias from variance selection.

### 3.2 Rank selection with signal

Rank selection is the procedure for selecting the $\phi_i$ with the largest value (or the top $K$ $\phi_i$'s with the largest values). It is the simplest selection policy and the one that we are instinctively most likely to use. We have seen previously how rank selection can introduce significant bias. In the bio-marker example, suppose there is *no signal* in the data, then $X_{i,j} \sim \mathcal{N}(0,1)$ and $\phi_i \sim \mathcal{N}(0, 1/n)$. Under rank selection, $\phi_T$ would have a bias close to $\sqrt{(2\log m)/n}$.

A basic question is what is the bias of rank selection when there *is* signal in the data. Our framework cleanly illustrates how signal in the data can reduce rank selection bias. As before, this insight follows transparently from studying the mutual information $I(T, \boldsymbol{\phi})$. Recall that mutual information is bounded by entropy: $I(T;\boldsymbol{\phi}) \leq H(T) \leq \log(m)$. When the data provides a strong signal of which $T$ to select, the distribution of $T$ is far from uniform, and $H(T)$ is much smaller than its worst case value of $\log(m)$.

Consider the following simple example. Assume

$$\phi_i \sim \begin{cases} N(\mu, \sigma^2) & \text{If } i = I^* \\ N(0, \sigma^2) & \text{If } i \neq I^* \end{cases}$$

where $\mu \geq 0$. The data analyst would like to identify $I^*$ and report the value of $T^*$. To do this, she selects $T = \arg\max_i \phi_i$. When $\mu = 0$, there is no true signal in the data and $T$ is equally likely to take on any value in $\{1, .., m\}$, $I(T;\boldsymbol{\phi}) = H(T) = \log(m)$. As $\mu$ increases, however, $T$ concentrates on $I^*$, causing $H(T)$ and the bias $\mathbf{E}[\phi_T - \mu_T]$ to diminish. This is reflected in Fig.1.

### 3.3 Regularization via randomized selection

The previous section illustrates how signal in the data intrinsically reduces selection bias by reducing the $H(T)$ term in $I(T;\boldsymbol{\phi}) = H(T) - H(T|\boldsymbol{\phi})$. A complementary approach to reduce bias is to increase $H(T|\boldsymbol{\phi})$ by adding *randomization* to the selection policy $T$. It's easy to maximize $H(T|\boldsymbol{\phi})$ by choosing $T$ uniformly at random from $\{1, ..., m\}$, independently of $\boldsymbol{\phi}$. Imagine however that we want to not only ensure that $H(T|\boldsymbol{\phi})$ is large, but want to choose $T$ such that $\phi_T$ is large. After observing $\boldsymbol{\phi}$, it's natural then to set the probability $\pi_i$ of setting $T = i$ by solving a maximization problem

$$\begin{aligned} \underset{\boldsymbol{\pi}}{\text{maximize}} \quad & H(\boldsymbol{\pi}) \\ \text{subject to} \quad & \sum_{i=1}^{k} \pi_i \phi_i \geq b \text{ and } \sum_{i=1}^{k} \pi_i = 1. \end{aligned}$$

The solution $\boldsymbol{\pi}^*$ to this problem is the maximum entropy or "Gibbs" distribution, which sets

$$\pi_i^* \propto e^{-\beta \phi_i} \qquad i \in \{1, .., m\} \tag{1}$$

for $\beta > 0$ that is chosen so that $\sum_i \pi_i^* \phi_i = b$. This procedure effectively adds stability, or a kind of regularization, to the selection strategy by adding randomization. Whereas tiny perturbations to $\boldsymbol{\phi}$ may change the identify of $T = \arg\max_i \phi_i$, the distribution $\pi^*$ is relatively insensitive to small changes in $\boldsymbol{\phi}$. Note that the strategy (1) is one of the most widely studied algorithms in the field of online learning [5], where it is often called *exponential weights*. In our framework it is transparent how it reduces bias.

To illustrate the effect of randomized selection, we use simulations to explore the tradeoff between bias and accuracy. We consider the following simple randomization scheme:

1. Take as input parameters $\beta$ and $K$, and observations $\phi_1, ...\phi_m$. Here $\beta$ is the inverse temperature in the Gibbs distribution and $K$ is number of $\phi_i$'s we need to select.

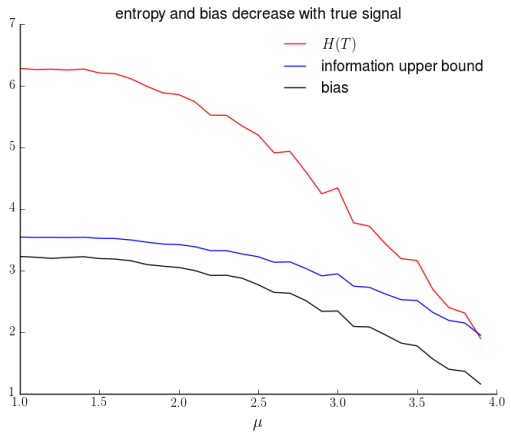


Figure 1: As the signal strength increases ($\mu$ increases), the entropy of selection $H(T)$ decreases, causing the information upper bound to also decrease. The bias of the selection decreases as well.

2. Sample without replacement $K$ indices $T_1, ...T_K$ from $\boldsymbol{\pi}^*$ given in (1). Report the corresponding values $\phi_{T_1}, ..., \phi_{T_k}$.

We consider settings where we have two groups of $\phi_i$'s: after relabeling assume that $\mu_1 = ... = \mu_{N_1} = \mu > 0$ and $\mu_i = 0$ for $i > N_1$. We define the *bias* of the selection to be $\frac{1}{K}\sum_i(\phi_{T_i} - \mu_i)$ and the *accuracy* of the selection to be $|\{T_i : T_i < N_1\}|/K$. In Fig.2, we illustrate the tradeoff between accuracy and bias for $N_1 = 1000, n - N_1 = 100000$ (i.e. there are many more false signals than true signals), randomization strength $\beta = 2$, and the signal strength $\mu$ varying from 1 to 5. As predicted, randomized selection significantly decreased bias. In the low signal regime ($\mu \leq 1$), both rank selection and randomized selection have low accuracy because the signal is overwhelmed by the large number of false positives. In the high signal regime ($\mu \geq 4$), both selection methods have accuracy close to one and rank selection has significantly less bias. In the intermediate regime ($1 < \mu < 4$), randomized selection has substantially less bias but is less accurate.

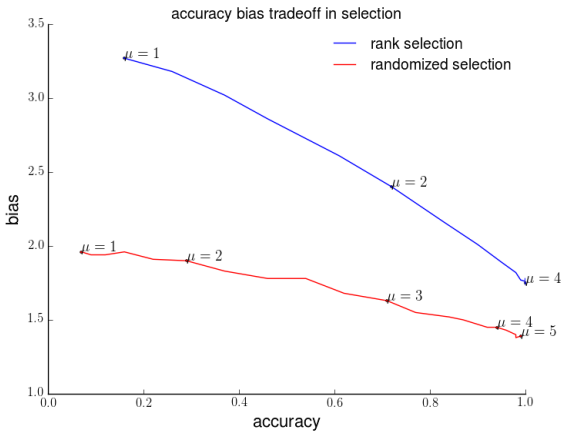


Figure 2: Tradeoff between accuracy and bias as the signal strength $\mu$ increases. The two curves illustrate the tradeoff for the Gibbs randomized selection procedure and the standard rank selection procedure of selecting the top $K = 100$ with the largest $\phi_i$'s.

# 4 Controlling bias with information budget

Given a well-defined selection protocol, the last section illustrates how we can reason about and quantify the bias of the protocol via its mutual information. In many settings of data analysis, the analysis process is much more adaptive, possibly involving many rounds of adaptation. This *research degrees of freedom* is ubiquitous and it is often not feasible to model it by one concise protocol. This takes us beyond the standard setting of selective inference. In this section, we show how our mutual information framework can be straightforwardly used to analyze bias under a general model of adaptive data analysis, allowing for arbitrary researcher degrees of freedom. Our approach naturally suggests the notion of controlling bias by having a limited budget of mutual information for the entire analysis process. We discuss the close connections to recent work inspired by differential privacy [3, 7, 8, 11].

## 4.1 General model of adaptive data analysis

We consider a general model of adaptive data analysis similar to that of Dwork et al. [7, 8].

1. At step 1, the analyst selects a hypothesis $\phi_{T_1}$ to query for $T_1 \in [m]$ and receives a response $Y_{T_1} \in \mathbb{R}$.
2. In the $k$-th iteration, the analyst chooses a hypothesis $\phi_{T_k}$ as a function of the results that she has received so far, $\{Y_{T_1}, T_1, ..., Y_{T_{k-1}}, T_{k-1}\}$, and receives feedback $Y_{T_k}$.
3. After $K$ iterations, the analyst selects $\phi_T \equiv \phi_{T_{K+1}}$ as a function of $\{Y_{T_1}, T_1, ..., Y_{T_K}, T_K\}$

The selection protocols from the Selective Inference section can be placed into this framework. Take rank selection for example, at the $k$-th step, $\phi_k$ is queried (i.e. the order is fixed and does not depend on the previous results) and $Y_k = \phi_k$ is returned. The analyst queries all $m$ $\phi_i$'s and returns the one with the max value. In general, $Y_{T_k}$ can differ from $\phi_{Y_k}$ and the number of iterations $K$ can be much less than $m$, which can be arbitrarily large. This general model of data analysis is consistent with multiple scenarios. For example, the data, $D$, could belong to some warehouse and the analyst can not access the $D$ directly. She makes queries $\phi_{T_k}$ and the warehouse returns answers $Y_{T_k}$. The analyst could also have the data $D$ herself, but implemented a system to return $Y_{T_k} = \phi_{T_k} +$ noise in order to reduce overfitting.

To apply our general result, Prop. 1, we want to exploit the structure of the adaptive analysis model to decompose $I(T_{k+1}; \boldsymbol{\phi})$. We prove the following *composition* lemma for mutual information.

**Lemma 1.** *Let* $H_k = (T_1, Y_{T_1}, T_2, Y_{T_2}, ..., T_k, Y_{T_k})$ *denote the history of interaction up to time* $k$. *Then, under the adaptive analysis model*

$$I(T_{k+1}; \boldsymbol{\phi}) \leq I(H_k; \boldsymbol{\phi}) = \sum_{i=1}^{k} I(Y_{T_i}; \phi_{T_i} | H_{i-1}, T_i)$$

The important takeaway from this Lemma is that by bounding the conditional mutual in-

formation between the response and the query, $I(Y_{T_i}; \phi_{T_i} | H_{i-1}, T_i)$, we can bound $I(T_{k+1}; \phi)$ and hence bound the bias after $k$ rounds of adaptive queries. Given a dataset $D$, we can imagine the analyst having a (mutual) *information budget*, $I^*$, which is decided a priori based on the size of the data and her tolerance for bias. At each step of the adaptive data analysis, how the analyst decides which hypothesis to test next (as a function of her analysis history) incurs an information cost quantified by $I(Y_{T_i}; \phi_{T_i} | H_{i-1}, T_i)$. The information costs accumulate additively over the analysis steps, until it reaches $I^*$, at which point the guarantee on bias requires the analysis to stop.

A trivial way to reduce mutual information is to return a response $Y_{T_i}$ that is independent of the query $\phi_{T_i}$, in which case the analyst learns nothing about the data and incurs no bias. However in order for the data to be useful for the analyst, we would like the results of the queries to also be accurate.

As before let $\mu_i = \mathbf{E}[\phi_i]$ denote the true answer of hypothesis $\phi_i$. We say that the query $\phi_i$ is answered with accuracy $(\tau, \epsilon)$ if $\mathbf{P}(|Y_i - \mu_i| \geq \tau) < \epsilon$. A natural goal would be to answer as many adaptive queries as possible for given $\tau, \epsilon$. We analyze a stylized case of the general model to show that we can answer $n^{2-\delta}$ queries accurately, where $n$ is the size of the dataset and $\delta$ is any constant larger than 0.

**Gaussian noise protocol.** We analyze the following special case.

1. Suppose $\phi_i \sim N(\mu_i, \frac{1}{n})$ and $\phi_1, ..., \phi_k$ is jointly Gaussian for any $k$.
2. For each query $\phi_{T_i}$, the protocol returns a distorted response $Y_{T_i} = \phi_{T_i} + W_{T_i}$ where $W_i \sim N(0, \frac{\omega}{n})$. Note that unlike $(\phi_1, \phi_2, ....)$, the sequence $(W_1, W_2, ....)$ is independent.

We want to know, as we scale $n \to \infty$, how many queries can be accurately answered as a function of $n$ by choosing the distortion level $\omega(n)$.

**Lemma 2.** *If $X \sim N(0, \sigma_1^2)$ and $Y = X + W$ where $W \sim N(0, \sigma_2^2)$ is independent of $X$, then*

$$I(X;Y) = \frac{1}{2} \log (1 + \beta) \leq \frac{\beta}{2}$$

*where $\beta = \sigma_1^2 / \sigma_2^2$ is the signal to noise ratio.*

**Proposition 7.** *When $\phi_i$'s are jointly Gaussian, the Gaussian noise protocol can answer $O(N^{2-\delta})$ queries accurately for any $\delta > 0$ and any accuracy parameters $(\tau, \epsilon)$.*

### 4.2 Connections to differential privacy

Adaptive analysis can be viewed as a Markov chain

$$T \leftarrow Y \equiv \{Y_{T_1}, ..., Y_{T_k}\} \leftarrow D \rightarrow \Phi \equiv \{\phi_1, ..., \phi_m\}.$$

By the information processing inequality [6], $I(T; \Phi) \leq I(Y; D) \leq H(Y)$. This motivates two families of strategies for controlling the mutual information $I(T; \Phi)$ in adaptive data analysis

1. ensure that the mutual information between the returned information and the underlying dataset, $I(Y; D)$, is small.
2. ensure that the total number of bits of information returned to the user, $H(Y)$, is small.

The first approach is similar to differential privacy and the second is similar to controlling the description length of the responses.

**Differential privacy** Let $\mathcal{A} : D \to \mathcal{Y}$ be an $\epsilon$-differentially private algorithm, where $D$ is a dataset of size $n$. Then $I_\infty(\mathcal{A}(D); D) \leq \ln \epsilon n$, where $I_\infty(Y; X) \equiv \log \max_x \frac{\mathbb{P}[Y=x]}{\mathbb{P}[X=x]}$ [8]. Note that $I(T; \Phi) \leq I(Y; X) \leq I_\infty(\mathcal{A}(X); X) \leq \ln \epsilon n$. Therefore if the selection policy $T$ corresponds to an $\epsilon$-differentially private mechanism, then the selection bias is bounded by $\sigma\sqrt{2 \ln \epsilon n}$. Differential privacy controls the max-information, $I_\infty$, which upper bounds our mutual information. Our mutual information bound can be used to quantify bias for general protocols that are not differentially private, e.g. rank selection.

**Minimal feedback** The recently proposed Ladder Mechanism [3] is an example of the second strategy, of explicitly minimizing the total amount of information in the feedback. Briefly, the Ladder Mechanism works as follows: at the $i$th step, $f_i : D \to [0, 1]$ is queried. Here $\phi_i = \frac{1}{n} \sum f_i(x_j)$ is the sample mean of $f_i$. If $\phi_i$ is the minimum among $\phi_1, ..., \phi_i$, then $y_i = \phi_i$ discretized into steps of size $\eta$ is returned. Otherwise, $y_i =$ NONE. Here $|\mathcal{Y}| = 1/\eta$. The number of possible distinct trajectories for $(y_1, ..., y_k)$ is $(k/\eta)^{1/\eta}$ and $H(Y) \leq \frac{1}{\eta} \log \frac{k}{\eta}$ grows logarithmic in $k$.

## 5 Discussion

We have introduced a general information theoretic approach to quantifying bias in adaptive data analysis. This conceptual framework lends insight into when existing analysis procedures lead to severe bias and when they do not. It also suggests engineering approaches to designing new analysis protocols with guaranteed low bias. An interesting direction of future work is to explore implementations of this approach in practical analytic settings.

## References


[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[2] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[3] Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *ICML 2015*, 2015.

[4] V Buldygin and K Moskvichova. The sub-gaussian norm of a binary random variable. *Theory of Probability and Mathematical Statistics*, 86:33–49, 2013.

[5] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[6] T.M. Cover and J.A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[7] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *arXiv preprint arXiv:1506.02629*, 2015.

[8] 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.

[9] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

[10] R.M. Gray. *Entropy and information theory*. Springer, 2011.

[11] Marcus Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 454–463. IEEE, 2014.

[12] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[13] P Kraft. Curses, winner's and otherwise, in genetic epidemiology. *Epidemiology*, pages 649–51, 2008.

[14] I Lee, G Lushington, and M Visvanathan. A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *Journal of Clinical Bioinformatics*, 2011.

[15] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[16] David McAllester. A pac-bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307.2118*, 2013.

[17] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, page 0956797611417632, 2011.

[18] Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

[19] Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani. Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 2014.

[20] Sara Van de Geer, Peter Buhlmann, Yaacov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[21] James Zou, Christoph Lippert, David Heckerman, Martin Aryee, and Jennifer Listgarten. Epigenome-wide association studies without the need for cell-type composition. *Nature Methods*, pages 309–11, 2014.