# How much does your data exploration overfit? Controlling bias via information usage.

Daniel Russo and James Zou



***Abstract*—Modern data is messy and high-dimensional, and it is often not clear a priori what are the right questions to ask. Instead, the analyst typically needs to use the data to search for interesting analyses to perform and hypotheses to test. This is an adaptive process, where the choice of analysis to be performed next depends on the results of the previous analyses on the same data. Ultimately, which results are reported can be heavily influenced by the data. It is widely recognized that this process, even if well-intentioned, can lead to biases and false discoveries, contributing to the crisis of reproducibility in science. But while any data-exploration renders standard statistical theory invalid, experience suggests that different types of exploratory analysis can lead to disparate levels of bias, and the degree of bias also depends on the particulars of the data set. In this paper, we propose a general information usage framework to quantify and provably bound the bias and other error metrics of an arbitrary exploratory analysis. We prove that our mutual information based bound is tight in natural settings, and then use it to give rigorous insights into when commonly used procedures do or do not lead to substantially biased estimation. Through the lens of information usage, we analyze the bias of specific exploration procedures such as filtering, rank selection and clustering. Our general framework also naturally motivates randomization techniques that provably reduce exploration bias while preserving the utility of the data analysis. We discuss the connections between our approach and related ideas from differential privacy and blinded data analysis, and supplement our results with illustrative simulations.**

***Index Terms*—Adaptive data analysis; Data snooping; Mutual information; Over-fitting;**

## I. Introduction

MODERN data is messy and high dimensional, and it is often not clear a priori what is the right analysis to perform. To extract the most insight, the analyst typically needs to perform exploratory analysis to make sense of the data and identify interesting hypotheses. This is invariably an adaptive process; patterns in the data observed in the first stages of analysis inform which tests are run next and the process iterates. Ultimately, the data itself may influence which results the analyst chooses to report, introducing *researcher degrees of freedom*: an additional source of over-fitting that isn't accounted for in reported statistical estimates [1]. Even if the analyst is well-intentioned, this exploration can lead to false discovery or large bias in reported estimates.

The practice of data-exploration is largely outside the domain of classical statistical theory. Standard tools of multiple hypothesis testing and false discovery rate (FDR) control assume that all the hypotheses to be tested, and the procedure for testing them, are chosen independently of the dataset. Any "peeking" at the data before committing to an analysis procedure renders classical statistical theory invalid. Nevertheless, data exploration is ubiquitous, and folklore and experience suggest the risk of false discoveries differs substantially depending on how the analyst explores the data. This creates a glaring gap between the messy practice of data analysis, and the standard theoretical frameworks used to understand statistical procedures. In this paper, we aim to narrow this gap. We develop a general framework based on the concept of information usage and systematically study the degree of bias introduced by different forms of exploratory analysis, in which the choice of which function of the data to report is made *after* observing and analyzing the dataset.

To concretely illustrate the challenges of data exploration, consider two data scientists Alice and Bob.

**Example 1.** *Alice has a dataset of 1000 individuals for a weight-loss biomarker study. For each individual, she has their weight measured at 3 time points and the current expression values of 2000 genes assayed from blood samples. There are three possible weight changes that Alice could have looked at—the difference between time points 1 and 2, 2 and 3 or 1 and 3—but Alice decides ahead of time to only analyze the weight change between 1 and 3. She computes the correlation across individuals between the expression of each gene and the weight change, and reports the gene with the highest correlations along with its $r^2$ value. This is a canonical setting where we have tools for controlling error in multiple-hypothesis testing and the false-discovery rate (FDR). It is well-recognized that even if the reported gene passes the multiple-testing threshold, its correlation in independent replication studies tend to be smaller than the reported correlation in the current study. This phenomenon is also called the Winner's Curse selection bias.*

**Example 2.** *Bob has the same data, and he performs some simple data exploration. He first uses data visualization to investigate the average expression of all the genes across all the individuals at each of the time points, and observes that there is very little difference between time 1 and 2 and there is a large jump between time 2 and 3 in the average expression. So he decides to focus on these later two time points. Next, he realizes that half of the genes always have low expression values and decides to simply filter them out. Finally, he computes the correlations between the expression of the 1000 post-filtered genes and the weight change between time 2 and 3. He selects the gene with the largest correlation and reports its value. Bob's analysis consists of three steps and the*

Daniel Russo is with the division of Decision, Risk and Operations at Columbia University.

James Zou is with Biomedical Data Science and, by courtesy, of Computer Science and Electrical Engineering at Stanford University

Manuscript received January 30, 2017; revised May 30, 2018; accepted September 12, 2019

Part of this work was presented at AISTATS 2016.

*results of each step depend on the results and decisions made in the previous steps. This adaptivity in Bob's exploration makes it difficult to apply standard statistical frameworks. We suspect there is also a selection bias here leading to the reported correlation being systematically larger than the real correlations if those genes are tested again. How do we think about and quantify the selection bias and overfitting due to this more complex data exploration? When is it larger or smaller than Alice's selection bias?*

The toy examples of Alice and Bob illustrate several subtleties of bias due to data exploration. First, the adaptivity of Bob's analysis makes it more difficult to quantify its bias compared to Alice's analysis. Second, for the same analysis procedure, the amount of selection bias depends on the dataset. Take Alice for example, if across the population one gene is substantially more correlated with weight change than all other genes, then we expect the magnitude of Winner's Curse decreases. Third, different steps of data exploration introduce different amounts of selection bias. Intuitively, Bob's visualizing of aggregate expression values in the beginning should not introduce as much selection bias as his selection of the top gene at the last step.

This paper introduces a mathematical framework to formalize these intuitions and to study selection bias from data exploration. The main tool we develop is a metric of the *bad* information usage in the data exploration. The true signal in a dataset is the signal that is preserved in a replication dataset, and the noise is what changes across different replications. Using Shannon's mutual information, we quantify the degree of dependence between the noise in the data and the choice of which result is reported. We then prove that the bias of an arbitrary data-exploration process is bounded by this measure of its bad information usage. This bound provides a quantitative measure of researcher degrees of freedom, and offers a single lens through which we investigate different forms of exploration.

In Section II, we present a general model of exploratory data-analysis that encompasses the procedures used by Alice and Bob. Then we define information usage and show how it upper and lower bounds various measures of bias and estimation error due to data exploration in Section IV. In Section V, we study specific examples of data exploration through the lens of information usage, which gives insight into Bob's practices of filtering, visualization, and maximum selection. Information usage naturally motivates randomization approaches to reduce bias and we explore this in Section VI. In Section VI, we also study a model of a data analyst who–like Bob–interacts adaptively with the data many times before selecting values to report.

## II. A Model of Data Exploration

We consider a general framework in which a dataset $D$ is drawn from a probability distribution $\mathcal{P}$ over a set of possible datasets $\mathcal{D}$. The analyst is considering a large number $m$ of possible analyses on the data, but wants to report only the most interesting results. She decides to report the result of a single analysis, and chooses which one *after* observing the realized dataset, $D$, or some summary statistics of $D$. More formally, the data analyst considers $m$ functions $\phi_1, ..., \phi_m : \mathcal{D} \to \mathbb{R}$ of the data, where $\phi_i(D)$ denotes the output of the $i$th analysis on the realization $D$. Each function $\phi_i$ is typically called an estimator; each $\phi_i(D)$ is an estimate or statistic calculated from the sampled data, and is a random variable due to the randomness in the realization of $D$. After observing the sampled-data, the analyst chooses to report the value $\phi_{T(D)}(D)$ for $T(D) \in \{1, ..., m\}$. 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 made by $T$ is itself a function of the sampled-data, the reported value $\phi_{T(D)}(D)$ may be significantly biased. For example, $\mathbf{E}[\phi_{T(D)}(D)]$ could be very far from zero even if each fixed function $\phi_i(D)$ has zero mean.

Note that although the number of estimators is assumed to be finite, it could be arbitrarily large; in particular $m$ can be exponential in the number of samples in the dataset. The $\phi_i$'s represent the set of all estimators that the analyst *potentially* could have considered during the course of exploration. Also, while for simplicity we focus on the case where exactly one estimate is selected and reported, our results apply in settings where the analyst selects and reports many estimates.[1]

**Example 1.** *For Alice, $D$ is a 1000-by-2003 matrix, where the rows are the individuals and the columns are the 2000 genes plus the three possible weight changes. Here there are $m = 2000$ potential estimators and $\phi_i$ is the correlation between the $i$th gene and the weight change between times 1 and 3. Alice's analysis corresponds to the selection procedure $T = \arg\max_i \phi_i$.*

**Example 2.** *Bob has the same dataset $D$. Because his exploration could have led him to use any of the three possible weight-change measures, the set of potential estimators are the correlations between the expression of one gene and one of the three weight changes and there are $2000 \times 3$ such $\phi_i$'s. Bob's adaptive exploration also corresponds to a selection procedure $T$ that takes the dataset and picks out a particular correlation value $\phi_T$ to report.*

**Selection Bias.** Denote the *true* value of estimator $\phi_i$ as $\mu_i \equiv \mathbf{E}[\phi_i(D)]$; this is the value that we expect if we apply $\phi_i$ on multiple independent replication datasets. On a particular dataset $D$, if $T(D) = i$ is the selected test, the output of data exploration is the value $\phi_i(D)$. The output and true-value can be written more concisely as $\phi_T$ and $\mu_T$. The difference $\phi_T - \mu_T$ captures the error in the reported value. We are interested in quantifying the *bias* due to data-exploration, which is defined as the average error $\mathbf{E}[\phi_T - \mu_T]$. We will quantify other metrics of error, such as the expected absolute-error $\mathbf{E}[|\phi_T - \mu_T|]$ or the squared-error $\mathbf{E}[(\phi_T - \mu_T)^2]$. In each case, the expectation is over all the randomness in the dataset $D$ and any intrinsic randomness in $T$.

## III. Related Work

There is a large body of work on methods for providing meaningful statistical inference and preventing false discov-

[1]For example, if the analyst chooses to report $m_0 \leq m$ results, our framework can be used to bound the average bias of the reported values by letting $T$ be a random draw from the $m_0$ selected analyses.

ery. Much of this literature has focused on controlling the false discovery rate in multiple-hypothesis testing where the hypotheses are not adaptively chosen [2, 3]. Another line of work studies confidence intervals and significance tests for parameter estimates in sparse high dimensional linear regression (see [4, 5, 6, 7] and the references therein).

One recent line of work [8, 9] 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 correct for selection bias by explicitly conditioning on the event that a particular model was chosen. While some powerful results can be derived in the selective inference framework (e.g. [10, 11]), 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 before looking at the data. Our approach does not explicitly adjust for selection bias, but it enables us to formalize insights that apply to very general selection procedures. For example, the selection rule $T$ could represent the choice made by a data-analyst, like Bob, after performing several rounds of exploratory analysis.

A powerful line of work in computer science and learning theory [12, 13, 14] has explored the role of algorithmic stability in preventing overfitting. Related to stability is PAC-Bayes analysis, which provides powerful generalization bounds in terms of KL-divergence [15]. There are two key differences between stability and our framework of information usage. First, stability is typically defined in the worst case setting and is agnostic of the data distribution. An algorithm is stable if, no matter the data distribution, changing one training point does not affect the predictions too much. Information usage gives more fine-grained bias bounds that depend on the data distribution. For example, in Section V-C we show the same learning algorithm has lower bias and lower information usage as the signal in the data increases. The second difference is that stability analysis has been traditionally applied to prediction problems—i.e. to bounding generalization loss in prediction tasks. Information usage applies to prediction—e.g. $\phi_i$ could be the squared loss of a classifier—but it also applies to model estimation where $\phi_i$ could be the value of the $i$th parameter.

Exciting recent work in computer science [16, 17, 18, 19] has leveraged the connection between algorithmic stability and differential privacy to design specific differentially private mechanisms that reduce bias in adaptive data analysis. In this framework, the data analyst interacts with a dataset indirectly, and sees only the noisy output of a differentially private mechanism. In Section VI, we discuss how information usage also motivates using various forms of randomization to reduce bias. In the Appendix, we discuss the connections between mutual information and a recently introduced measure called max-information [19]. The results from this privacy 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 not differentially private. The results in algorithmic stability and differential privacy are complementary to our framework: these approaches are specific techniques that guarantee low bias for worst-case analysts, while our framework quantifies the bias of any general data-analyst.

Finally it is also important to note the various practical approaches used in specific settings to quantify or reduce bias from exploration. Using random subsets of data for validation is a common prescription against overfitting. This is feasible if the data points are independent and identically distributed samples. However, for structured data—e.g. time-series or network data—it is not clear how to create a validation set. The bounds on overfitting we derive based on information usage do not assume independence and apply to structured data. Special cases of selection procedures $T$ corresponding to filtering by summary statistics of biomarkers [20] and selection matrix factorization based on a stability criterion [21] have been studied. The insights from these specific settings agree with our general result that low information usage limits selection bias.

## IV. Controlling Exploration Bias via Information Usage

**Information usage upper bounds bias.** In this paper, we bound the degree of bias in terms of an information–theoretic quantity: the mutual information between the choice $T(D)$ of which estimate to report, and the actual realized value of the estimates $(\phi_1(D), ..., \phi_m(D))$. 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.** *If $\phi_i - \mu_i$ is $\sigma$–sub-Gaussian for each $i \in \{1, ..., m\}$, then,*

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

*where $I$ denotes mutual information*[2].

The randomness of $\boldsymbol{\phi}$ is due to the randomness in the realization of the data $D \sim \mathcal{P}$. This captures how each estimate $\phi_i$ varies if a replication dataset is collected, and hence captures the *noise* in the statistics. The mutual information $I(T; \boldsymbol{\phi})$, which we call **information usage**, then quantifies the dependence of the selection process on the noise in the estimates. Intuitively, a selection process that is more sensitive to the noise (high $I$) is at a greater risk for bias. We will also refer to $I(T; \boldsymbol{\phi})$ as bad information usage to highlight the intuition that it really captures how much information about the noise in the data goes into selecting which estimate to report. We normally think of data analysis as trying to extract the *good* information, i.e. the true signal, from data. The more bad information is used, the more likely the analysis procedure is to overfit.

[2]The mutual information between two random variables $X, Y$ is defined as $I(X; Y) = \sum_{x,y} \mathbf{P}(x, y) \log \left( \frac{\mathbf{P}(x,y)}{\mathbf{P}(x)\mathbf{P}(y)} \right)$.

When $T$ is determined entirely from the values $\{\phi_1, ..., \phi_m\}$, mutual information $I(T; \boldsymbol{\phi})$ is equal to entropy $H(T)$. This quantifies how much $T$ varies over different independent replications of the data.

The parameter $\sigma$ provides the natural scaling for the values of $\phi_i$. The condition that $\phi_i$ is $\sigma$-sub-Gaussian ensures that its tail is not too heavy[3]. In the Appendix, we show how this condition can be relaxed to treat cases where $\phi_i$ is a sub-Exponential random variables (Proposition 9) as well as settings where the $\phi_i$'s have different scaling $\sigma_i$'s (Proposition 8).

Proposition 1 applies in a very general setting. The magnitude of overfitting depends on the generating distribution of the data set, and on the size of the data, and this is all implicitly captured in by the mutual-information $I(T; \boldsymbol{\phi})$. For example, a common type of estimate of interest is $\phi_i = n^{-1} \sum_{j=1}^{n} f_i(X_j)$, 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

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

To illustrate Proposition 1, we consider two extreme settings: one where $T$ is chosen independently of the data and one where $T$ heavily depends on the values of all the $\phi_i$'s. The subsequent sections will investigate the applications of information usage in depth in settings that interpolate between these two extremes.

**Example: data-agnostic exploration.** Suppose $T$ is independent of $\boldsymbol{\phi}$. This may happen if the choice of which estimate to report is decided ahead of time and cannot change based on the actual data. It may also occur when the dataset can be split into two statistically independent parts, and separate parts are reserved for data-exploration and estimation. In such cases, one expects there is no bias because the selection does not depend on the actual values of the estimates. This is reflected in our bound: since $T$ is independent of $\boldsymbol{\phi}$, $I(T; \boldsymbol{\phi}) = 0$ and therefore $\mathbf{E}[\phi_T] = \mathbf{E}[\mu_T]$.

**Example: maximum of Gaussians.** Suppose each $\phi_i$ is an independent sample from the zero-mean normal $\mathcal{N}(0, \sigma^2)$. If $T = \arg\max_{1 \leq i \leq m} \phi_i$, then $I(T; \boldsymbol{\phi}) = H(T) = \log(m)$ because all $m$ $\phi_i$'s are symmetric and have equal chance of being selected by $T$. Applying Proposition 1 gives $\mathbf{E}[\phi_T - \mu_T] = \mathbf{E}[\phi_T] \leq \sigma\sqrt{2\log(m)}$. This is the well known inequality for the maximum of Gaussian random variables. Moreover, it is also known that this equation approaches equality as the number of Gaussians, $m$, increases, implying that the information usage $I(T; \boldsymbol{\phi})$ precisely measures the bias of max-selection in this setting. It is illustrative to also consider a more general selection $T$ which first ranks the $\phi_i$'s from the largest to the smallest and then uniformly randomly selects one of the $m_0$ largest $\phi_i$'s to report. Here $I(T; \boldsymbol{\phi}) = H(T) - H(T|\boldsymbol{\phi})$, where $H(T) = \log m$ (by the symmetry of $\phi_i$ as before) and $H(T|\boldsymbol{\phi}) = \log m_0$ (since given the values of $\phi_i$'s there is still uniform randomness over which of the top $m_0$ is selected). We immediately have the following corollary.

**Corollary 1.** *Suppose for each $i \in \{1, ..., m\}$, $\phi_i$ is a zero-centered sub-Gaussian random variable with parameter $\sigma$. Let $\phi_{(1)} \geq \phi_{(2)} \geq ... \geq \phi_{(m)}$ denote the values of $\phi_i$ sorted from the largest to the smallest. Then*

$$\mathbf{E}\left[\frac{1}{m_0}\sum_{i=1}^{m_0} \phi_{(i)}\right] \leq \sigma\sqrt{2\log\frac{m}{m_0}}.$$

In Appendix C, we show that this bound is also tight as $m$ and $m_0$ increase.

**Information usage bounds other metrics of exploration error.** So far we have discussed how mutual information upper bounds the bias $|\mathbf{E}[\phi_T - \mu_T]|$. In different application settings, it might be useful to control other measures of exploration error, such as the absolute error deviation $\mathbf{E}[|\phi_T - \mu_T|]$ and the squared error $\mathbf{E}\left[(\phi_T - \mu_T)^2\right]$.

Here we extend Proposition 1 and show how $\sqrt{I(T; \boldsymbol{\phi})}$ and $I(T; \boldsymbol{\phi})$ can be used to bound absolute error deviation and squared error. Note that due to inherent noise even in the absence of selection bias, the absolute or squared error can be of order $\sigma$ or $\sigma^2$, respectively. The next result effectively bounds the additional error introduced by data-exploration in terms of information-usage.

**Proposition 2.** *Suppose for each $i \in \{1, ..., m\}$, $\phi_i - \mu_i$ is $\sigma$ sub-Gaussian. Then*

$$\mathbf{E}[|\phi_T - \mu_T|] \leq \sigma + c_1\sigma\sqrt{2I(T; \boldsymbol{\phi})}$$

*and*

$$\mathbf{E}[(\phi_T - \mu_T)^2] \leq 1.25\sigma^2 + c_2\sigma^2 I(T; \boldsymbol{\phi}).$$

*where $c_1 < 36$ and $c_2 \leq 10$ are universal constants.*

**Information usage also lower bounds error.** In the maximum of Gaussians example, we have already seen a setting where information usage precisely quantifies bias. Here we show that this is a more general phenomenon by exhibiting a much broader setting in which mutual-information lower bounds expected-error. This complements the upper bounds of Proposition 1 and Proposition 2.

Suppose $T = \arg\max_i \phi_i$ where $\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}, I)$. Because $T$ is a deterministic function of $\boldsymbol{\phi}$, mutual information is equal to entropy. The probability $T = i$ is a complicated function of the mean vector $\boldsymbol{\mu}$, and the entropy $H(T)$ provides a single number measuring the uncertainty in the selection process. Proposition 2 upper bounds the average squared distance between $\phi_T$ and $\mu_T$ by entropy. The next proposition provides a matching lower bound, and therefore establishes a fundamental link between information usage and selection-risk in a natural family of models.

**Proposition 3.** *Let $T = \arg\max_{1 \leq i \leq m} \phi_i$ where $\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}, I)$. There exist universal numerical constants $c_1 = 1/8$, $c_2 < 2.5$ , $c_3 = 10$, and $c_4 = 1.5$ such that for any $m \in \mathbb{N}$ and $\boldsymbol{\mu} \in \mathbb{R}^m$,*

$$c_1 H(T) - c_2 \leq \mathbf{E}[(\phi_T - \mu_T)^2] \leq c_3 H(T) + c_4.$$

[3] A random variable $X$ is said to be $\sigma$-sub-Gaussian if $\mathbf{E}\left[e^{\lambda(X - \mathbf{E}[X])}\right] \leq e^{\sigma^2\lambda^2/2}$ for all $\lambda$.

Recall that the entropy of $T$ is defined as

$$H(T) = \sum_i \mathbf{P}(T=i) \log\left(\frac{1}{\mathbf{P}(T=i)}\right).$$

Here $\log(1/\mathbf{P}(T=i))$ is often interpreted as the "surprise" associated with the event $\{T=i\}$ and entropy is interpreted as expected surprise in the realization of $T$. Proposition 3 relies on a link between the surprise associated with the selection of statistic $i$, and the squared error $(\phi_i - \mu_i)^2$ on events when it is selected.

To understand this result, it is instructive to instead consider a simpler setting; imagine $m = 2$ , $\phi_1 = x$ always, $\phi_2 \sim \mathcal{N}(0,1)$, and the selection rule is $T = \arg\max_i \phi_i$. When $x >> 0$ is large,

$$\log(1/\mathbf{P}(T=2)) = \log(1/\mathbf{P}(\phi_2 \geq x)) \approx x^2/2$$

and so the surprise associated with the event $\{T=2\}$ scales with the squared gap between the selection threshold $x$ and the true mean of $\phi_2$. One can show that as $x \to \infty$,

$$\begin{aligned} H(T_x) &\sim \mathbf{P}(T_x=2)\log(1/\mathbf{P}(T_x=2)) \\ &\sim \mathbf{P}(T_x=2)x^2 \\ &\sim \mathbf{E}[(\phi_{T_x} - \mu_{T_x})^2] \end{aligned}$$

where $T_x$ denotes the selection rule with threshold $x$ and $f(x) \sim g(x)$ if $f(x)/g(x) \to 1$ as $x \to \infty$.

In the Appendix, we investigate additional threshold-based selection policies applied to Gaussian and exponential random variables, allowing for arbitrary correlation among the $\phi_i$'s, and show that $H(T)$ also provides a natural lower bound on estimation-error.

## V. When is bias large or small? The view from information usage

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. Information usage provides a unified framework to understand selection bias in these settings. Our results inform when these these procedures introduce significant selection bias and when they do not. 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 value of $\log(m)$. We provide several simulation experiments as illustrations.

### A. Filtering by marginal statistics

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

$$\begin{aligned} I(T;\boldsymbol{\phi}) &= H(T) - H(T|\boldsymbol{\phi}) \\ &\leq H(T) - I(T;D|\boldsymbol{\phi}) \\ &= (1-\alpha)H(T) \end{aligned}$$

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 [22, 23]. 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. Many studies first perform a filtering step to select only the markers that have high variance and remove the rest. The rationale 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 $T = \arg\max_i V_i$ where $V_i = \sum_{j=1}^n (X_{i,j} - \phi_i)^2$. Here we consider the case where only the marker with the largest variance is selected, but all the discussion applies to softer selection when we select the $K$ 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. In fact, when the $X_{i,j}$ 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.

This illustrates an important point that the bias bound depends on $I(T;\boldsymbol{\phi})$ instead of $I(T;D)$. The selection process $T$ may depend heavily on the dataset $D$ and $I(T;D)$ could be large. However as long as the statistics of the data used for selection have low mutual information with the estimators $\phi_i$, there is low bias on the reported values.

We can apply our framework to analyze biases that arise from feature filtering more generally. A common practice in data analysis is to reduce multiple hypotheses testing burden and increase discovery power by first filtering out covariates or features that are unlikely to be relevant or interesting [20]. This can be viewed as a two-step procedure. For each feature $i$, two marginal statistics are computed from the data, $\psi_i$ and $\phi_i$. Filtering corresponds to a selection protocol on $\psi_i$. Since $I(T;\boldsymbol{\phi}) \leq I(\boldsymbol{\psi};\boldsymbol{\phi})$, if the $\psi_i$'s do not reveal too much information about $\phi_i$'s then the filtering step does not create too much bias. In our example above, $\psi_i$ is the sample variance and $\phi_i$ is the sample mean of feature $i$. General principles for creating independent $\psi_i$ and $\phi_i$ are given in [20].

More generally, suppose the dataset determines two sets of statistics $\boldsymbol{\phi} = (\phi_1, .., \phi_m)$ and $\boldsymbol{\psi} = (\psi_i, ..., \psi_{m'})$. We report $\phi_T$ and want to quantify its bias, but the selection rule depends only on the $\psi_i$'s, i.e. $T = f(\psi_i)$ can be expressed as a function of the $\psi_i$'s. This captures the general situation where data processing and feature selection uses one set of summary statistics ($\psi$) and we want to quantify the bias introduced in these steps on another set of statistics ($\phi$). The dependence structure can be expressed as a Markov chain $T - \boldsymbol{\psi} - \boldsymbol{\phi}$, where this notation indicates that conditioned on $\boldsymbol{\psi}$, $T$ is independent of $\boldsymbol{\phi}$. The data processing inequality

implies $I(T;\boldsymbol{\phi}) \leq I(\boldsymbol{\phi};\boldsymbol{\psi})$, which–combined with our bound–formalizes the intuition that the selection rule cannot be substantially biased when $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ share limited information in common. However, this bound may be quite loose. We instead turn to strong data processing inequalities.

**Definition 1.** *A pair of random variables $(X,Y)$ satisfies a strong data-processing inequality with contraction coefficient $\eta \in [0,1]$ if for all random variables $U$ with $U - X - Y$*

$$I(U;Y) \leq \eta I(U;X)$$

*Let $\eta_{XY}$ be the smallest constant such that* (1) *is satisfied for all valid $U$.*

The contraction coefficient satisfies several natural properties. First, it *tensorizes* [24]. That is, if $(X_1,Y_1),...(X_n,Y_n)$ is an independent sequence, then $\eta_{XY} = \max_i \eta_{X_iY_i}$. Also, if $X,Y$ and $Z$ follow a Markov chain $X - Y - Z$ then $\eta_{XZ} \leq \eta_{YZ}$.

**Example.** Suppose $D = (X_1,...,X_n)$ consists of $n$ iid random variables and $\boldsymbol{\psi} = (X_1,...,X_k)$ is a subsample of $k < n$ data points. Then $\eta_{\boldsymbol{\psi}\boldsymbol{\phi}} \leq \eta_{\boldsymbol{\psi}D} \leq k/n$ [25].

**Example.**(Noisy Channels) If $(X,Y)$ corresponds to a binary symmetric channel with error rate $\delta$ then $\eta_{XY} = (1-2\delta)^2$ [26].

Note that the contraction coefficient $\eta_{\boldsymbol{\phi}\boldsymbol{\psi}}$ depends only on the distribution of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, and not on the selection rule $T$. A benefit of our mutual information framework for bounding the exploration bias is that we can immediately apply Strong Data Processing to obtain tighter bounds on bias:

**Proposition 4.** *Suppose $\phi_i - \mu_i$ is $\sigma$ sub-Gaussian for each $i \in \{1,..,,m\}$. Then if the selection $T$ is independent of $\boldsymbol{\phi}$ conditioned on $\boldsymbol{\psi}$,*

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

### B. Bias due to data visualization

Data visualization, using clustering for example, is a common technique to explore data and it can inform subsequent analysis. How much selection bias can be introduced by such visualization? While in principle a visualization could reveal details about every data point, a human analyst typically only extracts certain salient features from plots. For concreteness, we use clustering as an example, and imagine the analyst extracts the number of clusters $K$ from the analysis. In our framework the natural object of study is the information usage $I(K;\boldsymbol{\phi})$, since if the final selection $T$ is a function of $K$, then $I(T;\boldsymbol{\phi}) \leq I(K;\boldsymbol{\phi})$ by the data-processing inequality. In general, $K$ is a random variable that can take on values 1 to $n$ (if each point is assigned its own cluster). When there is structure in the data and the clustering algorithm captures it, then $K$ can be strongly concentrated around a specific number of clusters and $I(K;\boldsymbol{\phi}) \leq H(K) \approx 0$. In this setting, clustering is informative to the analyst but does not lead to "bad information-usage" and therefore does not increase exploration bias. This is a stylized example; if the analyst uses additional information beyond the number of clusters $K$, then the bias could increase.

### C. 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 in Subsection V-A, suppose there is *no signal* in the data, so $X_{i,j} \sim \mathcal{N}(0,1)$ and $\phi_i \sim \mathcal{N}(0,1/n)$. Under rank selection, $\phi_T$ would have a bias close to $\sqrt{(2\log m)/n}$.

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

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

where $\mu \geq 0$. The data analyst would like to identify $I^*$ and report the value of $\phi_{I^*}$. 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. We simulated this example with $m = 1000$ $\phi_i$'s, all but one of which are i.i.d. samples from $\mathcal{N}(0,1)$ and $\phi_{I^*} \sim \mathcal{N}(\mu,1)$ for $\mu \in [1,4]$. The simulation results, averaged over 1000 independent runs, are shown in Figure 1.

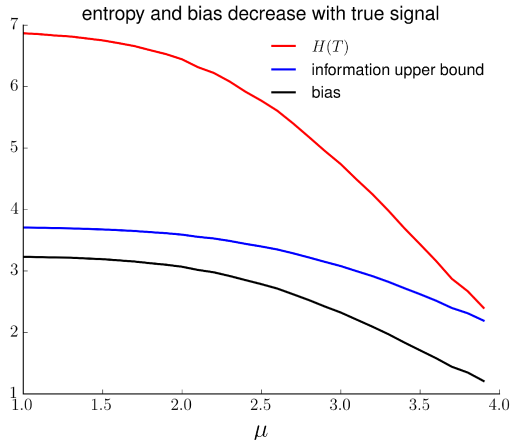


Figure 1. As the signal strength increases ($\mu$ increases), the entropy of selection $H(T)$ decreases, causing the information upper bound $\sqrt{2I(T;\boldsymbol{\phi})}$ to also decrease. The bias of the selected $\phi_T$ decreases as well.

### D. Information usage along the Least Angle Regression path

Our analyses illustrate that in certain stylized settings, information usage tightly bounds the bias of optimization

selections. Here we show that information usage also accurately captures the bias of a more complex selection procedure corresponding to Least Angle Regressions (LARS) [27]. LARS is an interesting example for two reasons. First it is widely used as a practical tool for sparse regression and is closely related to LASSO. Second LARS composes a sequence of maximum selections and thus provides a more complex example of selection. In Figure 2, we show the simulation results for LARS under three data settings corresponding to low, medium and high signal-to-noise ratios. We use bootstrapping to empirically estimate the information usage and since we know the ground truth of the experiment, we can easily compute the bias of LARS. As the signal in the data increases, the information usage of LARS decreases and, consistent with the predictions of our theory, the bias of LARS also decreases. Moreover, as the number of selected features increases, the average (per feature) information usage of LARS decreases and, consistent with this, the average bias of LARS also decreases monotonically. Details of the experiment are in the Appendix.

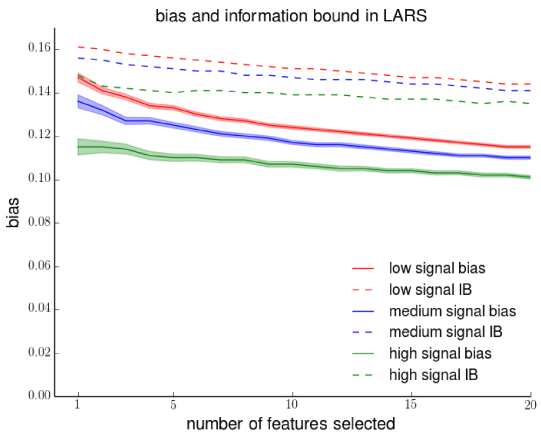


Figure 2. Information bound $\sqrt{2I(T;\phi)}$ (dotted lines) and bias of Least Angle Regression (solid lines). Results are shown for low (red), medium (blue) and high (green) signal-to-noise settings. The $x$-axis indicates the number of features selected by LARS and the $y$-axis corresponds to the average information usage and bias in the selected features.

### *E. Differentially private algorithms*

Recent papers [28, 19] have shown that techniques from differential privacy, which were initially inspired by the need to protect the security and privacy of datasets, can be used to develop adaptive data analysis algorithms with provable bounds on over-fitting. These differentially private algorithms satisfy worst case bounds on certain likelihood ratios, and are guaranteed to have low information-usage. On the other hand, many algorithms have low information-usage without being differentially private. Moreover, as we have seen, the exploration bias of an algorithm could be large or small depending on the particular dataset (e.g. the signal-to-noise ratio of the data) and information usage captures this. Differentially private algorithms have low information usage for *all* datasets and $T$ that is designed adversarial to exploit this dataset, so this is a much stricter condition. In [19], the authors also define and study a notion of max-information, which can be viewed as a worst-case analogue of mutual information. We discuss the relationship between these measures further in the Appendix.

### *F. Information usage and classification overfitting*

This section applies our framework to the problem of overfitting in classification. A classifier is trained on a dataset consisting of $n$ examples, with input features $X_1, .., X_n \in \mathcal{X}$ and corresponding labels $Y_1, ...Y_n \in \{-1, 1\}$. We consider here a setting where the features of the training examples $X_i = x_i$ are fixed, and study overfitting of the noisy labels. Each label $Y_i$ is drawn independently of the other labels from an unknown distribution $\mathbf{P}(Y_i = 1|X_i = x_i)$. A classifier $f$ associates a label $f(x) \in \{-1, 1\}$ with each input $x$. The training error of a fixed classifier $f$ is

$$\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(f(x_i) \neq Y_i)$$

while its true error rate is

$$L(f) = \mathbf{E}[\hat{L}(f)] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{P}(f(x_i) \neq Y_i),$$

is the expected fraction of examples it mis-classifies on a random draw of the labels $Y_1, .., Y_n$. The process of training a classifier corresponds to selecting, as a function of the observed data, a particular classification rule $\hat{f}$ from a large family $\mathcal{F}$ of possible rules. Such a procedure may overfit the training data, causing the average training error $\mathbf{E}[\hat{L}(\hat{f})]$ to be much smaller than its true error rate $\mathbf{E}[L(\hat{f})]$.

As an example, suppose each $X_i \in \mathbb{R}^d$ is a $d$–dimensional feature vector, and $\mathcal{F} = \{f_\theta : \theta \in \mathbb{R}^d\}$ consists of all linear classifiers of the form $f_\theta(x) = \mathbf{1}(x^T\theta \geq 0)$. A training algorithm might set $\hat{f} = f_{\hat{\theta}}$ by choosing the parameter vector that minimizes the number of mis-classifications on the training set. This procedure tends to overfit the noise in the training data, and as a result the average training of $\hat{f}$ can be much smaller than its true error rate. The risk of over-fitting tends to increase with the dimension $d$, since higher dimensional models allow the algorithm to fit more complicated, but spurious, patterns in the training set.

The field of statistical learning provides numerous bounds on the magnitude of overfitting based on more general notions of the complexity of an arbitrary function class $\mathcal{F}$, with the most influential being the Vapnik-Chervonenkis dimension, or VC-dimension[4]. While the focus is on overfitting of the training data, similar concerns apply to overfitting the validation data.

The next proposition provides information-usage bounds the degree of over-fitting, and then shows that mutual information is upper-bounded by the VC–dimension of $\mathcal{F}$. Therefore, information-usage is always constrained by function-class complexity.

[4]The VC-dimension of $\mathcal{F}$ is the size of the largest set it shatters. A set $\{x_1, .., x_m\} \in \mathcal{X}$ is shattered by $\mathcal{F}$ if for any choice of labels $y_1, .., y_m \in \mathcal{Y}$, there is some $f \in \mathcal{F}$ with $f(x_i) = y_i$ for all $i$.

**Proposition 5.** *Let* $\mathbf{x} \equiv (x_1, ..., x_n)$, $\mathbf{Y} \equiv (Y_1, ..., Y_n)$, $\hat{f}(\mathbf{x}) \equiv (\hat{f}(x_1), ...\hat{f}(x_n))$ *and* $\log_+(z) \equiv \max\{1, \log(z)\}$. *Then,*

$$\mathbf{E}[L(\hat{f}) - \hat{L}(\hat{f})] \leq \sqrt{\frac{I(\hat{f}(\mathbf{x}); \mathbf{Y})}{2n}}.$$

*If* $\mathcal{F}$ *has VC-dimension* $d < \infty$, *then*

$$I(\hat{f}(\mathbf{x}); \mathbf{Y}) \leq d \log_+ \left(\frac{ne}{d}\right).$$

The proof of the information usage bound follows by an easy reduction to Proposition 1. The proof of the second claim relies on a known link between VC-dimension and a notion of the log-covering numbers of the function-class.

It is worth highlighting that because VC-dimension depends only on the class of functions $\mathcal{F}$, bounds based on this measure can't shed light on which types of data-generating distributions and fitting procedures $(\mathbf{X}, \mathbf{Y}) \mapsto \hat{f}$ allow for effective generalization. Information usage depends on both, and a result could be much smaller than VC-dimension; for example, this occurs when some classifiers in $\mathcal{F}$ are much more likely to be selected after training than others. This can occur naturally due to properties of the training procedure, like regularization, or properties of the data-generating distribution.

### G. Approximately independent data splitting.

A data scientist has access to data in the form of $n$ samples $(s_1, \ldots, s_n)$ from a Markov chain. She would like to mimic the honest data-splitting she uses with i.i.d data. To do this, she splits the into three parts: $(s_1, .., s_{n_1})$, $(s_{n_1+1}, .., s_{n_2})$ and $(s_{n_2+1}, .., s_n)$. The first part is used for selection, the third for estimation, and the middle data is thrown away. In particular, $\phi = \phi_1, ..., \phi_m : (s_{n_2+1}, .., s_n) \to \mathbb{R}^m$ and that $T : (s_1, .., s_{n_1}) \mapsto \{1, \ldots, m\}$. One expects that if $n_2 - n_1$ is large so there is a sufficient delay between the two samples, then the risk of bias and overfitting will be low. We'll see that this is easy to formalize via an information usage lens.

We assume the Markov process is stationary and time homogeneous with stationary distribution $\pi$. Moreover, it satisfies a uniform mixing condition

$$\max_s D(\mathbf{P}(s_\tau = \cdot | s_1 = s) \,||\, \pi) \leq c_0 e^{-c_1 \tau} \qquad \forall \tau \in \mathbb{N}.$$

We then claim that

$$I(T; \boldsymbol{\phi}) \leq c_0 e^{-c_1(n_2 - n_1)}$$

and so a sufficient delay between the sample used for selection and the sample used for estimation guarantees low bias. We have immediately that,

$$\begin{aligned} I(s_{t+\tau}; s_t) &= \sum_s \mathbf{P}(s_t = s) D(\mathbf{P}(s_{t+\tau} = \cdot | s_t = s) \,||\, \mathbf{P}(s_t = \cdot)) \\ &\leq c_0 e^{-c_1 \tau} \end{aligned}$$

where we used that $\mathbf{P}(s_t = s) = \pi(s)$. Then, by the data processing inequality

$$\begin{aligned} I(T; \boldsymbol{\phi}) &\leq I((s_1, .., s_{n_1}); (s_{n_2+1}, .., s_n)) \\ &\leq I(s_{n_1}; s_{n_2+1}) \\ &\leq c_0 e^{-c_1(n_2 - n_1)}. \end{aligned}$$

### H. Bias control via FDR control

There has been intense interest in large-scale hypothesis testing procedures that control the false-discovery rate. Here we consider the bias and error incurred when estimation is performed after variables are selected in this manner, and bound this in terms of the false discovery rate and the rates of type I and type II errors.

As motivation, consider analysis of a large micro-array experiment. There is a large set of gene-expression data $D \in \mathbb{R}^{n \times m}$ consisting of $m$ gene expression levels drawn from $n$ samples, where there first $n_1$ samples were taken from tissue with a cancerous tumor and the remaining $n_2 = n - n_1$ were taken from healthy tissue. A scientist would like to identify genes with large differential between the expression levels across the two tissue types. She casts this as a multiple hypothesis testing problem, where rejecting a given null hypothesis indicates strength of evidence that an observed differential is unlikely due to random chance. Many procedures exist to control the false discovery rate, which is the expected proportion of type I errors among rejected null hypotheses.

Consider for example the procedure proposed by Benjamini and Hochberg. One first constructs p-values $p_1, \ldots p_m$ for $m$ separate hypothesis testing problems. These are then sorted as $p_{(1)} \leq p_{(2)} \leq \ldots, p_{(m)}$. To guarantee the false discovery rate is controlled at some level $q \in (0, 1)$, their procedure specifies the selection of the the first $\hat{t}$ hypotheses, where $\hat{t}$ is the largest number such that $p_{(t)} \leq qt/m$. Framed differently, all hypotheses with p-values less than a random threshold $\hat{l} = q\hat{t}/m$ are rejected. To gain some insight, let us consider a simple model where each $p$-value is drawn either from a uniform distribution (i.e. the null distribution) or an alternative distribution $F$. Consider an asymptotic regime where the number of alternative $m \to \infty$, but the proportion of alternatives following the null distribution stays fixed. Then [29] show that under regularity conditions on $F$, the random threshold $\hat{l}$ converges in probability to a deterministic limit $l^*$. Therefore, the rate of type I and type II errors, as well as the proportion of false discoveries, all tend to a fixed levels asymptotically as $m \to \infty$. Whether a particular hypothesis is accepted or rejected is still random and data-dependent, but when $m$ is large the overall proportions are nearly deterministic.

We consider a more abstract framework. There is some random matrix $D \in \mathbb{R}^{n \times m}$, and a vector $\phi \in \mathbb{R}^m$ that is a function of $D$ with $\boldsymbol{\mu} = \mathbf{E}[\boldsymbol{\phi}]$. The indices $\{1, \ldots, m\}$ are partitioned into two sets $\mathcal{H}_0$ and $\mathcal{H}_1$. A selection procedure is a map $\psi : \mathbb{R}^{n \times m} \to \{0, 1\}^m$, where $\psi(D)_i = 1$ indicates variable $i$ was selected. We set $S_1 \subset \{1, \ldots, m\}$ to be the set of selected variables and $S_0$ to be its complement.

To form the analogy with the story above, we think of $\phi$ as a vector of summary statistics of the columns of $D$—e.g. the observed gene expression differential between tumor tissue and healthy tissue—and think of $\mathcal{H}_0$ as the set for which the null distribution holds — e.g. across repeated samples there would not be an observed differential. The selected variables $S_1$ is the set for which the null hypothesis was rejected. Set $\hat{\alpha} = \#(\mathcal{H}_0 \cap S_1)/\#\mathcal{H}_0$ and $\hat{\beta} = \#(\mathcal{H}_1 \cap S_0)/\#\mathcal{H}_1$ to be analogues of the proportion of type I and type II errors. Note

that $\hat{\alpha}$ is the fraction of false discoveries relative to the total number of nulls, and is different from what is called the False Discovery Proportion or FDP. To simplify the discussion, we assume there is always at least one selected variable, so $S_1$ is nonempty. We are interested in the average error or bias in reported estimates among selected, which leads to the study of quantities like

$$\frac{1}{\#S_1}\sum_{i\in S_1}(\phi_i-\mu_i), \quad \frac{1}{\#S_1}\sum_{i\in S_1}|\phi_i-\mu_i|, \quad \text{or} \quad \frac{1}{\#S_1}\sum_{i\in S_1}(\phi_i-\mu_i)^2. \tag{1}$$

These can be rewritten as $\mathbf{E}[\phi_T-\mu_T]$, $\mathbf{E}[|\phi_T-\mu_T|]$ or $\mathbf{E}[(\phi_T-\mu_T)^2]$ where, conditioned on $D$, $T$ is drawn uniformly at random from the set of selected of selected variables $S_1$. This leads naturally to the study of information usage $I(T;\boldsymbol{\phi})$, which bounds these quantities. The quantities in (1) reflect whether, the estimation procedures applied to the selected variables produce accurate results *on average*. For this reason, we are able to provide meaningful guarantees that do not degrade as $m\to\infty$, a regime in which it is impossible to guarantees that *every* selected variable is estimated accurately.

Now, let us define $\mathrm{FDR}=\mathbf{P}(T\in\mathcal{H}_0)$ to be the false discovery rate. This is the expected proportion of selected variables $S_1$ that are contained within the null set $\mathcal{H}_0$. The next lemma bounds information usage in terms of the false discovery rate, the rates of type I and II error, and an extra error term that vanishes as the random proportion of realized type I and II errors concentrate around their expected value. A short proof is given in Appendix E.

**Proposition 6.** *For the FDR control problem defined above,*

$$I(T;\boldsymbol{\phi}) \le h(\mathrm{FDR}) + (1-\mathrm{FDR})\cdot\log\left(\frac{1}{1-\beta}\right) + \mathrm{FDR}\cdot\log\left(\frac{1}{\alpha}\right)+\xi$$

*where $h(p)=-p\log(p)-(1-p)\log(1-p)$ denotes the binary entropy function, $\alpha=\mathbf{E}[\hat{\alpha}]$ and $\beta=\mathbf{E}[\hat{\beta}]$ denote the type I and II error proportion relative to the total number of true null and true alternative, respectively. The error term is*

$$\xi=\mathbf{E}\left[\log_+\left(\frac{1-\beta}{1-\hat{\beta}}\right)\right]+\mathbf{E}\left[\log_+\left(\frac{\alpha}{\hat{\alpha}}\right)\right].$$

*for* $\log_+(x)\equiv\max\{0,\log(x)\}$.

This result further formalizes the insight that estimation after selection is unlikely to overfit in settings where the selection procedure works reliably. When the rates of false discovery, type I error, and type II error are small, information usage is guaranteed to also be low. The implied bounds on estimation error after selection grow smoothly as the reliability of the selection procedure degrades.

## VI. Limiting information usage and bias via randomization

We have seen how information usage provides a unified framework to investigate the magnitude of exploration bias across different analysis procedures and datasets. It also suggests that methods that reduces the mutual information between $T$ and $\boldsymbol{\phi}$ can reduce bias. In this section, we explore simple procedures that leverages randomization to reduce information usage and hence bias, while still preserving the utility of the data analysis.

We first revisit the rank-selection policy considered in the previous subsection, and derive a variant of this scheme that uses randomization to limit information usage. We then consider a model of a human data analyst who interacts sequentially with the data. We use a stylized model to show that, even if the analysts procedure is unknown or difficult to describe, adding noise during the data-exploration process can provably limit the bias incurred. Many authors have investigated adding noise as a technique to reduce selection bias in specialized settings [28, 30]. The main goal of this section is to illustrate how the effects of adding noise is transparent through the lens of information usage.

### A. Regularization via randomized selection

Subsection V-C illustrates how signal in the data intrinsically reduces the bias of rank selection by reducing the entropy term $H(T)$ in $I(T;\boldsymbol{\phi})=H(T)-H(T|\boldsymbol{\phi})$. A complementary approach to potentially reduce bias is to increase conditional entropy $H(T|\boldsymbol{\phi})$ by adding *randomization* to the selection policy $T$. Note that while this randomization increases $H(T|\boldsymbol{\phi})$, it also increases $H(T)$ and thus could increase information usage. It is easy to maximize conditional entropy by choosing $T$ uniformly at random from $\{1,...,m\}$, independently of $\boldsymbol{\phi}$. Imagine however that we want to not only ensure that conditional entropy is large, but want to choose $T$ such that the selected value $\phi_T$ is large. After observing $\boldsymbol{\phi}$, it is natural then to set the probability $\pi_i$ of setting $T=i$ by solving a maximization problem

$$\begin{aligned}\underset{\boldsymbol{\pi}\in\mathbb{R}_+^m}{\text{maximize}} \quad & H(\boldsymbol{\pi})\\ \text{subject to} \quad & \sum_{i=1}^k \pi_i\phi_i \ge b \text{ and } \sum_{i=1}^k \pi_i = 1.\end{aligned}$$

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

$$\pi_i^* \propto e^{\beta\phi_i} \qquad i\in\{1,..,m\} \tag{2}$$

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 identity of $T=\arg\max_i\phi_i$, the distribution $\pi^*$ is relatively insensitive to small changes in $\boldsymbol{\phi}$. Note that the strategy (2) is one of the most widely studied algorithms in the field of online learning [31], where it is often called *exponential weights*. It is also known as the exponential mechanism in differential privacy. 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, max-entropy randomization scheme:

- 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.
- Sample without replacement $K$ indices $T_1, ...T_K$ from $\boldsymbol{\pi}^*$ given in (2). 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 $\frac{1}{K}\sum_{i=1}^{K}(\phi_{T_i} - \mu_{T_i})$ and the *accuracy* of the selection to be $|\{T_i : T_i \leq N_1\}|/K$, which is the fraction of reported $\phi_{T_i}$ with true signal $\mu$. In Figure 3, 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. Consistent with the theoretical analysis, max-entropy selection significantly decreased bias. In the low signal regime ($\mu = 1$), both rank selection and max-entropy 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 max-entropy selection has significantly less bias. In the intermediate regime ($1 < \mu < 4$), max-entropy selection has substantially less bias but is less accurate than rank selection.

Formally, unless the Gibbs distributions is degenerate with probability 1,

$$I(T; \boldsymbol{\phi}) = H(T) - H(T|\boldsymbol{\phi}) \leq \log(m) - H(T|\boldsymbol{\phi}) < \log(m),$$

so information usage is strictly smaller than its worst-case value of $\log(m)$. It is worth highlighting, however, that the Gibbs mechanism described above does not reduce bias or information usage for all possible data–generating distributions because it could increase entropy $H(T)$.

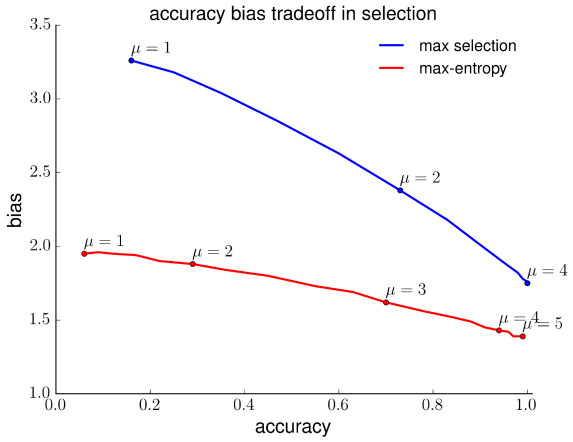


Figure 3. Tradeoff between accuracy and bias as the signal strength $\mu$ increases. The two curves illustrate the tradeoff for the maximum selection (i.e. reporting the largest $K = 100$ values of $\phi_i$) and the max-entropy randomized selection procedures.

### B. Randomization for a multi-step analyst

We next study how randomization can decrease information usage and bias even when we have very little knowledge of what the analyst is doing. To illustrate this idea, we analyze in detail a simple example of a very flexible data analyst who performs multiple steps of analysis. Flexibility in multi-step data analysis presents a challenge to current statistical approaches for quantifying selection bias. Recent development in post-selection inference have focused on settings where the selection rule is simple and analytically tractable, and the full analysis procedure is fixed and specified before any data analysis is performed. While powerful results can be derived in this framework—including exact bias corrections and valid post-selection confidence intervals [8, 9]—these methods do not apply for exploratory analysis where the procedure can be quite flexible.

In this section, we show how our mutual information framework can be used to analyze bias for a flexible multi-step analyst. We show that even if one does not know, or can't fully describe, the selection procedure $T$, one can control its bias by controlling the information it uses. The main idea is to inject a small amount of randomization at each step of the analysis. This randomization is guaranteed to keep the bad information usage low *no matter what the analyst does.*

The idea of adding randomization during data analysis to reduce overfitting has been implemented as practical rule-of-thumb in several communities. Particle physicists, for example, have advocated *blind data analysis*: when deciding which results to report, the analyst interacts with a dataset that has been obfuscated through various means, such as adding noise to observations, removing some data points, or switching data-labels. The raw, uncorrupted, dataset is only used in computing the final reported values [32]. Adding noise is also closely related to a recent line of work inspired by differential privacy [16, 18, 19, 17].

**A model of flexible, multi-step analyst.** We consider a model of adaptive data analysis similar to that of [19, 18]. In this setting, the analyst learns about the data by running a series of analyses on the dataset. Each analysis is modeled by a function of the data $\phi_i$, and choice of which analysis to run may depend on the results from all the earlier analyses. More formally, we define the model as follows:

1) At step 1, the analyst selects a statistic $\phi_{T_1}$ to query for $T_1 \in [m]$ and observes a result $Y_{T_1} \in \mathbb{R}$.
2) In the $k$-th iteration, the analyst chooses a statistic $\phi_{T_k}$ 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 result $Y_{T_k}$.
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 simplest setting is when the result of the analysis is just the value of $\phi_{T_k}$ on the data $D$: $Y_{T_K} = \phi_{T_K}(D)$. An example of this is the rank selection considered before. 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 maximal value.

In general, we allow the analysis output $Y_{T_K}$ to differ from the empirical value of the test $\phi_{T_K}$ and a particularly useful form is $Y_{T_k} = \phi_{T_k} +$ noise . This captures blind analysis settings, where the analyst intentionally adds noise throughout the data analysis in order to reduce over-fitting. A natural goal is to ensure that for every query $T_k$ used in the adaptive analysis, the reported result $Y_{T_K}$ is close to true value $\mu_{T_K}$. We will show through analyzing the information usage that noise addition can indeed guarantee such accuracy.

This adaptive analysis protocol can be viewed as a Markov chain

$$T_{k+1} \leftarrow H_k \equiv \{T_1, Y_{T_1}, ..., T_k, Y_{T_k}\} \leftarrow D \rightarrow \boldsymbol{\phi},$$

where recall that $\boldsymbol{\phi}$ denotes the vector $\{\phi_1, ..., \phi_m\}$. By the information processing inequality [33], $I(T_{k+1}; \boldsymbol{\phi}) \leq I(H_k; \boldsymbol{\phi})$. Therefore, a procedure that controls the mutual information between the history of feedback $H_k$ and the statistics $\boldsymbol{\phi}$ will automatically control the mutual information $I(T_{k+1}; \boldsymbol{\phi})$. By exploiting the structure of the adaptive analysis model, we can decompose the cumulative mutual information $I(H_k; \boldsymbol{\phi})$ into a sum of $k$ terms. This is formalized in 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 information between the response and the queried value at each step, $I(Y_{T_i}; \phi_{T_i} | H_{i-1}, T_i)$, 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*, $I_b$, 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, the analyst's choice of statistic to query 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 $I_b$, 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.

**Adding randomization to reduce bias.** As before let $\mu_i = \mathbf{E}[\phi_i]$ denote the true answer of query $\phi_i$. If each $\phi_i - \mu_i$ is $\sigma$–sub-Gaussian, then $\mathbf{E}[|\phi_i - \mu_i|] \leq \sigma$. Using Proposition 2, we can bound the average excess error of the response $Y_{T_k}$, $\mathbf{E}[|Y_{T_k} - \mu_{T_k}|] - \sigma$, by the sum of two terms,

$$\begin{aligned}
&\mathbf{E}[|Y_{T_k} - \mu_{T_k}|] - \sigma \\
&\leq \mathbf{E}[|Y_{T_k} - \phi_{T_k}|] + \mathbf{E}[|\phi_{T_k} - \mu_{T_k}| - \sigma] \\
&\leq \underbrace{\mathbf{E}[|Y_{T_k} - \phi_{T_k}|]}_{\text{Distortion}} + \underbrace{c\sigma\sqrt{2I(T_k; \boldsymbol{\phi})}}_{\text{Selection Bias}}.
\end{aligned}$$

Response accuracy degrades with distortion, a measure of the magnitude of the noise added to responses, but this distortion also controls the degree of selection bias in future rounds. We will explicitly analyze the tradeoff between these terms in a stylized case of the general model.

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

1) Suppose $\phi_i \sim \mathcal{N}(\mu_i, \frac{\sigma^2}{n})$ and $(\phi_1, ..., \phi_k)$ is jointly Gaussian for any $k$.
2) For the $j$th query $\phi_{T_j}$, $j = 1, 2, ...$, the protocol returns a distorted response $Y_{T_j} = \phi_{T_j} + W_j$ where $W_j \sim \mathcal{N}(0, \frac{\omega_j^2}{n})$. Note that unlike $(\phi_1, \phi_2, ....)$, the sequence $(W_1, W_2, ....)$ is independent.

The term $n$ can be thought of as the number of samples in the data-set. Indeed, if $\phi_i$ is the empirical average of $n$ samples from a $\mathcal{N}(\mu_i, \sigma^2)$ distribution, then $\phi_i \sim \mathcal{N}(\mu_i, \sigma^2/n)$. The ratio $\sigma^2/\omega_j^2$ is the signal-to-noise ratio of the $k$th response. We want to choose the distortion levels $(\omega_1, \omega_2, ...)$ so as to guarantee that a large number of queries can be answered accurately. In order to do this, we will use the next lemma to relate the distortion levels to the information provided by a response. The lemma gives a form for the mutual information $I(X; X + W)$ where $X$ and $W$ are independent Gaussian random variables. As one would expect, this shows that mutual information is very small when the variance of $W$ is much larger than the variance of $X$. Lemma 3, provided in the Appendix, provides a similar result when $X$ is a general (not necessarily Gaussian) random variable.

**Lemma 2.** *If $X \sim \mathcal{N}(0, \sigma_1^2)$ and $Y = X + W$ where $W \sim \mathcal{N}(0, \sigma_2^2)$ 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.*

Using Lemma 2, we provide an explicit bound on the accuracy of $Y_{T_{k+1}}$ as a function a function of $n, \sigma$ and $k$. Note that this result places no restriction on the procedure that generates $(T_1, T_2, ...)$ except that the choice $T_k$ can depend on $\boldsymbol{\phi}$ only through the data $\{T_1, Y_{T_1}, ...T_{k-1}, Y_{T_{k-1}}\}$ available at time $k$.

**Proposition 7.** *Suppose $\phi_i \sim \mathcal{N}(\mu_i, \frac{\sigma^2}{n})$ and $(\phi_1, ..., \phi_k)$ is jointly Gaussian for any $k$. If for the $j$th query, $Y_{T_j} = \phi_{T_j} + W_j$ where $W_j \sim \mathcal{N}(0, \frac{\sigma^2\sqrt{j}}{n})$ and $(W_1, W_2, ...)$ is independent of $\boldsymbol{\phi}$, then for every $k \in \mathbb{N}$*

$$\mathbf{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] \leq c\left(\frac{\sigma k^{1/4}}{n^{1/2}}\right)$$

*where $c$ denote a universal constant that is independent of $\sigma, \omega, k$, and $n$.*

If the sequence of choices $(T_1, T_2, T_3, ...)$ were non-adaptive, simply returning responses without any noise ($Y_{T_i} = \phi_{T_i}$) would guarantee $\mathbf{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] \leq \sigma/\sqrt{n}$. In the adaptive model, the first few queries are still answered with accuracy of order $\sigma/\sqrt{n}$, but the error increases for the later queries. This illustrates the fundamental tension that the longer

the analyst explores the data, the more likely for the later analysis to overfit.

The factor $k^{1/4}$ can roughly be viewed as the worst-case price of adaptivity. It is worth emphasizing this price would be more severe if the system returned responses without any noise. When no noise is added error can be as large as $\mathbf{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] = \Omega(\sigma\sqrt{k/n})$, as is demonstrated in Example 1 in the Appendix. Therefore, adding noise offers a fundamental improvement in attainable performance.

A similar insight was attained by [28], who noted that by adding Laplacian noise it is possible to answer up to $n^2$ queries accurately, whereas without noise accuracy degrades after $n$ queries. In the Gaussian case, it's clear from our bound that as $n, k \to \infty$, all queries will be answered accurately as long as $k = o(n^2)$.

## VII. Discussion

We have introduced a general information usage approach to quantify bias that arises from data exploration. While we focus on bias, we show our mutual information based metric can be used to bound other error metrics of interest, such as the average absolute error $\mathbf{E}[|\phi_T - \mu_T|]$. It is interesting to note that the same information usage also naturally appears in the lower bound on error, suggesting it may be fundamentally linked to exploration bias. This paper established lower bounds when the selection process corresponds to solving optimization problems—i.e. $T = \arg\max$. An interesting direction of research is to understand more general exploration procedures in which information usage provides a tight approximation to bias.

One advantage of using mutual information to bound bias is that we have many tools to analyze and compute mutual information. This conceptual framework allow us to extract insight into settings when common data analysis procedures lead to severe bias and when they do not. In particular we show how signal in the data can reduce selection bias. Information usage also suggests engineering approaches to reduce mutual information (and hence bias) by adding randomization to each step of the data exploration. Another important project is to investigate implementations of such randomization approaches in practical analytic settings.

As discussed before, the information usage framework proposed here is very much complementary to the exciting developments in post-selection inference and differential privacy. Post-selection inference, for very specific settings, is able to exactly characterize and correct for exploration biases—in this case exploration is feature and model selection. Differential privacy lies at the other extreme in that it derives powerful but potentially conservative results that apply to an adversarial data-analyst. The modern practice of data science often lies in between these two extremes—the analyst has more flexibility than assumed in post-selection inference, but is also interested in finding true signals and hence is much less adversarial than the worst-case. Information usage provides a bound on exploration bias in all settings. It is also important that this bound is data-dependent. In practice, the same analyst may be much less prone to false discoveries when exploring a high-signal dataset versus a low-signal dataset, and this should be reflected in the bias metric. An interesting goal is to develop approaches that combine the sharpness of post-selection inference and differential privacy with the generality of information usage.

## Appendix A
## Overview of the Appendix

The appendix provides complete proofs of all the results in the main text as well as extensions and additional applications of information usage. Section B gives the proof of Proposition 1, which states that information usage can be used to upper bounds selection bias. We also show that more general results hold when the estimators have different variances and when the estimators have heavier tales (i.e. sub-exponential rather than sub-Gaussian). Section C then proves that the error due to exploration is at least as large as the information usage for several families of explorations, which includes Proposition 3. Section D completes the proof of the link between information usage and classification overfitting (Proposition 5). In Section F, we provide additional applications to show how information usage can be used to control the bias in other metrics of interest, such as p-values in a multiple hypothesis testing problem and regret in optimization under uncertainty. Section G provides additional details of the experiments corresponding to Figure 2. Section H completes the analysis of how randomization controls the bias of a multi-step, flexibile data analyst. Section I discusses how our information usage relates to other information measures such as max-information.

## Appendix B
## Proofs of Information Usage Upper Bounds

### A. Information Usage Upper Bounds Bias: Proof of Proposition 1

The proof of Proposition 1 relies on the following variational form of Kullback–Leibler divergence, which is given in Theorem 5.2.1 of Robert Gray's textbook *Entropy and Information Theory* [34].

**Fact 1.** *Fix two probability measures* $\mathbf{P}$ *and* $\mathbf{Q}$ *defined on a common measurable space* $(\Omega, \mathcal{F})$. *Suppose that* $\mathbf{P}$ *is absolutely continuous with respect to* $\mathbf{Q}$. *Then*

$$D\left(\mathbf{P}||\mathbf{Q}\right) = \sup_X \left\{\mathbf{E}_{\mathbf{P}}[X] - \log \mathbf{E}_{\mathbf{Q}}[e^X]\right\},$$

*where the supremum is taken over all random variables* $X$ *such that the expectation of* $X$ *under* $\mathbf{P}$ *is well defined, and* $e^X$ *is integrable under* $\mathbf{Q}$.

*Proof of Proposition 1.*

$$\begin{aligned} I(T;\boldsymbol{\phi}) &= \sum_{i=1}^{n} \mathbf{P}(T=i) D\left(\mathbf{P}(\boldsymbol{\phi} = \cdot | T = i) \,||\, \mathbf{P}(\boldsymbol{\phi} = \cdot)\right) \\ &\geq \sum_{i=1}^{n} \mathbf{P}(T=i) D\left(\mathbf{P}(\phi_i = \cdot | T = i) \,||\, \mathbf{P}(\phi_i = \cdot)\right) \end{aligned}$$

Applying Fact 1 with $\mathbf{P} = \mathbf{P}(\phi_i = \cdot|T = i)$, $\mathbf{Q} = \mathbf{P}(\phi_i = \cdot)$, and $X = \lambda(\phi_i - \mu_i)$, we have

$$D\left(\mathbf{P}(\phi_i = \cdot|T = i)\,||\,\mathbf{P}(\phi_i = \cdot)\right) \geq \sup_{\lambda} \lambda\Delta_i - \lambda^2\sigma^2/2$$

where $\Delta_i \equiv \mathbf{E}[\phi_i|T = i] - \mu_i$. Taking the derivative with respect to $\lambda$, we find that the optimizer is $\lambda = \Delta_i/\sigma^2$. This gives

$$2\sigma^2 I(T;\boldsymbol{\phi}) \quad \geq \quad \sum_{i=1}^{n} \mathbf{P}(T = i)\Delta_i^2 = \mathbf{E}[\Delta_T^2].$$

By the tower property of conditional expectation and Jensen's inequality

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

□

**Remark.** *In the first step of the proof of Proposition 1, we used the fact that, for all $i \in \{1, ..., m\}$,*

$$\begin{aligned} & D\left(\mathbf{P}(\boldsymbol{\phi} = \cdot|T = i)\,||\,\mathbf{P}(\boldsymbol{\phi} = \cdot)\right) \\ \geq \quad & D\left(\mathbf{P}(\phi_i = \cdot|T = i)\,||\,\mathbf{P}(\phi_i = \cdot)\right), \end{aligned}$$

*which follows from the information processing inequality. The application of this inequality is not tight in general and can lead to gaps between the actual bias and our upper bound based on $I(T;\boldsymbol{\phi})$. Consider the following scenario. Suppose $T : \phi_1 \to [2, ..., m]$, i.e. $T$ is a deterministic function that uses the realized value of $\phi_1$ to decide which* other *$\phi_j$ to select. For example, imagine $\phi_1 \sim \text{Uniform}[0, 1]$ and $T$ is defined so that $T = 2$ if $\phi_1 \in [0, 1/(m-1)]$, $T = 3$ if $\phi_i \in [1/(m-1), 2/(m-1)]$, and so on. Here $T$ is deterministic, $I(T;\boldsymbol{\phi}) = \log m$, and this is manifested in $D\left(\mathbf{P}(\boldsymbol{\phi} = \cdot|T = i)\,||\,\mathbf{P}(\boldsymbol{\phi} = \cdot)\right) > 0$. However, if $\phi_j, j \neq 1$ is independent of each other $\phi_i$, then $D\left(\mathbf{P}(\phi_i = \cdot|T = i)\,||\,\mathbf{P}(\phi_i = \cdot)\right) = 0$ and the bias is also 0. The upper bound of Proposition 1 is tight in other settings; it is also useful in general because the mutual information $I(T;\boldsymbol{\phi})$ is amenable to analysis and explicit calculation. In cases where there is a gap, we may study $\mathbf{P}(T = i)D\left(\mathbf{P}(\phi_i = \cdot|T = i)\,||\,\mathbf{P}(\phi_i = \cdot)\right)$ directly.*

### *B. Extension to Unequal Variances*

We can prove a generalization of Proposition 1 for settings when the estimates $\phi_i$ have unequal variances.

**Proposition 8.** *Suppose that for each $i \in \{1, ..., m\}$, $\phi_i - \mu_i$ is $\sigma_i$–sub-Gaussian. Then,*

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

*where $I$ denotes mutual information.*

*Proof.* The first part of the proof is the same as that of Proposition 1. For each $i \in \{1, ..., m\}$,

$$D\left(\mathbf{P}(\phi_i = \cdot|T = i)\,||\,\mathbf{P}(\phi_i = \cdot)\right) \geq \sup_{\lambda} \lambda\Delta_i - \lambda^2\sigma_i^2/2$$

where $\Delta_i \equiv \mathbf{E}[\phi_i|T = i] - \mu_i$. The optimizer is $\lambda_i = \Delta_i/\sigma_i^2$. Rearranging the terms gives

$$\Delta_i \leq \sigma_i\sqrt{2D\left(\mathbf{P}(\phi_i = \cdot|T = i)\,||\,\mathbf{P}(\phi_i = \cdot)\right)}.$$

This implies

$$\begin{aligned} & \mathbf{E}[\Delta_T] \\ = \quad & \sum_i \Delta_i \mathbf{P}(T = i) \\ \leq \quad & \sum_i \sigma_i \mathbf{P}(T = i)\sqrt{2D\left(\mathbf{P}(\phi_i = \cdot|T = i)\,||\,\mathbf{P}(\phi_i = \cdot)\right)} \\ \leq \quad & \sqrt{\sum_i \sigma_i^2 \mathbf{P}(\phi_i = \cdot|T = i)} \\ & \times \sqrt{2\sum_i \mathbf{P}(\phi_i = \cdot|T = i)D\left(\mathbf{P}(\phi_i = \cdot|T = i)\,||\,\mathbf{P}(\phi_i = \cdot)\right)} \\ = \quad & \sqrt{\mathbf{E}[\sigma_T^2]}\sqrt{2I(T;\boldsymbol{\phi})}. \end{aligned}$$

where we have used Cauchy-Schwartz for the second inequality.

□

### *C. Extension to Sub-exponential Random Variables*

Recall that a random variable $X$ is sub-Gaussian with parameter $\sigma$ if $\mathbf{E}[e^{\lambda(X-\mathbf{E}[X])}] \leq e^{\lambda^2\sigma^2/2}$ for *all* real-values $\lambda$. While many random variables are sub-Gaussian, there are other important classes of random variables that are light tailed, but not quite sub-Gaussian. Here, we will show how our information-usage bounds extend to the larger class of sub-exponential random variables. We say that $X$ is sub-exponential with parameters $(\sigma, b)$ if $\mathbf{E}[e^{\lambda(X-\mathbf{E}[X])}] \leq e^{\lambda^2\sigma^2/2}$ whenever $|\lambda| < 1/b$. For example if $X \sim \chi_n^2$ follows a chi-squared distribution with $n \geq 1$ degrees of freedom, then it is sub-exponential with parameters $(2\sqrt{n}, 4)$.

**Proposition 9.** *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 $b < 1$, we also have*

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

*Proof.* Following the same analysis as in the sub-Gaussian setting (Prop. 1), we have

$$D\left(\mathbf{P}(\phi_i = \cdot|T = i)\,||\,\mathbf{P}(\phi_i = \cdot)\right) \geq \sup_{\lambda<1/b} \lambda\Delta_i - \lambda^2\sigma^2/2$$

The RHS is greater than the value from setting $\lambda = 1/b$. Therefore, we have

$$D\left(\mathbf{P}(\phi_i = \cdot|T = i)\,||\,\mathbf{P}(\phi_i = \cdot)\right) \geq \frac{\Delta_i}{b} - \frac{\sigma^2}{2b^2}.$$

Multiplying each side by $P(T = i)$ and summing over $i \in \{1, .., m\}$ gives

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

and hence

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

When $b < 1$, $\lambda = 1/\sqrt{b} < 1/b$ is also a feasible point. Putting in this value of $\lambda$ into the calculations above gives the second bound

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

□

### D. Extension to Other Metrics of Exploration Error

**Proposition 2 - Part (1).** *Suppose for each $i \in \{1,...,m\}$, $\phi_i - \mu_i$ is $\sigma$ sub-Gaussian. Then*

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

*where $c < 36$ is a universal constant.*

*Proof.* Let $U_i = \phi_i - \mu_i$ which is assumed to be $\sigma$ sub-Gaussian and let $\gamma_i = \mathbf{E}[|\phi_i - \mu_i|]$ and $Y_i = |U_i| - \gamma_i$. We show below that $Y_i$ is sub-Gaussian with parameter $c\sigma$ where $c \leq 36$. This implies the result, since by Proposition 1 and the data-processing inequality,

$$\mathbf{E}[|\phi_T - \mu_T| - \gamma_T] = \mathbf{E}[Y_T] \leq c\sigma\sqrt{2I(T;\boldsymbol{Y})} \leq \sqrt{2I(T;\boldsymbol{\phi})}.$$

Since $\gamma_i \leq \sigma$ for all $i$, $\gamma_T \leq \sigma$, and we have

$$\mathbf{E}[|\phi_T - \mu_T|] \leq \sigma + 36\sigma\sqrt{2I(T;\boldsymbol{\phi})}.$$

The remainder of the proof shows $Y \equiv |U| - \mathbf{E}[|U|]$ is sub-Gaussian whenever $U$ is sub-Gaussian. We use the following equivalent definition of a sub-Gaussian random variable.

*a) Fact 1.:* [35] Given a zero-mean random variable $Y$, Suppose there is a constant $c \geq 1$ and Gaussian random variable $Z \sim \mathcal{N}(0,\tau^2)$ such that

$$\mathbf{P}(|Y| \geq s) \leq c(\mathbf{P}(|Z| \geq s)) \text{ for all } s \geq 0.$$

Then $Y$ is sub-Gaussian with parameter $\sqrt{2}c\tau$.

*b) Fact 2.:* [35] Suppose $Y$ is a zero-mean sub-Gaussian random variable with parameter $\sigma$. Then

$$\mathbf{P}(|Y| \geq s) \leq \sqrt{8}e\mathbf{P}(|Z| \geq s)$$

where $Z \sim \mathcal{N}(0, 2\sigma^2)$.

Let $U$ be a zero-mean random variable that is sub-Gaussian with parameter $\sigma$. Let $\gamma \equiv \mathbf{E}[|U|]$ and $Y \equiv |U| - \gamma$. We want to determine the sub-Gaussian parameter of $Y$. We have

$$\begin{aligned}\mathbf{P}(|Y| \geq s) &= \mathbf{P}(|U| \geq s + \gamma) + \mathbf{P}(|U| \leq \gamma - s) \\ &\leq \sqrt{8}e\mathbf{P}(|Z| \geq s) + \mathbf{P}(|U| \leq \gamma - s)\end{aligned}$$

where $Z \sim \mathcal{N}(0, 2\sigma^2)$ and we have used Fact 2. Moreover

$$\mathbf{P}(|U| \leq \gamma - s) \leq \frac{\mathbf{P}(|Z| \geq s)}{\mathbf{P}(|Z| \geq \gamma)}$$

since the RHS exceeds 1 for $s \leq \gamma$ and the LHS is 0 for $s > \gamma$. Hence

$$\mathbf{P}(|Y| \geq s) \leq \left(\sqrt{8}e + \frac{1}{\mathbf{P}(|Z| \geq \gamma)}\right)\mathbf{P}(|Z| \geq s)$$

and, by Fact 1, $Y$ is sub-Gaussian with parameter $2(\sqrt{8}e + 1/\mathbf{P}(|Z| \geq \gamma))\sigma$. We can simplify this expression further. Since $U$ is $\sigma$ sub-Gaussian, its variance is bounded above by $\sigma^2$. Therefore $\gamma \leq \sqrt{\mathbf{E}[U^2]} \leq \sigma$, which implies

$$\mathbf{P}(|Z| \geq \gamma) > \mathbf{P}(|Z| \geq \sigma) > 0.1$$

and $Y$ is sub-Gaussian with parameter $36\sigma$. □

This bound is similar to a bias-variance decomposition, where the $\sigma$ term is the variance and the mutual–information term is the bias. When selection is over many $\phi_i$'s, the bias term tends to dominate. The parameter $\sigma$ captures the magnitude of noise in the estimates, and therefore implicitly captures the number of samples in the data set. In particular, If $\phi_i = n^{-1}\sum_{j=1}^n f_i(X_j)$ where $\{f_i(X_j)\}_{j=1}^n$ is an independent sequence of $\sigma$-sub-Gaussian random variables, then

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

Using the fact that the square of a sub-Gaussian random variable is sub-exponential and Proposition 9, we can also control the mean squared distance between $\phi_T$ and $\mu_T$.

**Proposition 2 - Part (2).** *Suppose $\phi_i - \mu_i$ is $\sigma$ sub-Gaussian for each $i \in \{1,...,m\}$. Then*

$$\mathbf{E}[(\phi_T - \mu_T)^2] \leq \sigma^2\left(1.25 + 10I(T;\boldsymbol{\phi})\right).$$

*Proof.* We use the following fact about sub-Gaussian random variables.

*c) Fact 3.:* [35] If $Y$ be a zero-mean sub-Gaussian variable with parameter $\sigma$, then

$$\mathbf{E}\left[e^{\frac{\lambda Y^2}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}} \text{ for all } \lambda \in [0,1).$$

Given such a $Y$, we would like to derive the sub-exponential parameters of $Y^2 - \gamma$, where $\gamma \equiv \mathbf{E}[Y^2] \geq 0$. Applying Fact 3, we have

$$\mathbf{E}\left[e^{\frac{\lambda(Y^2-\gamma)}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}} \leq e^{10\lambda^2} \text{ for } \lambda \in [0, 0.1)$$

where the last inequality can be verified numerically. Using the substitution $t \equiv \lambda/\sigma^2$, we have

$$\mathbf{E}\left[e^{t(Y^2-\gamma)}\right] \leq e^{10\sigma^4 t^2} \text{ for } t \in \left[0, \frac{0.1}{\sigma^2}\right)$$

which implies that $Y^2 - \gamma$ is sub-exponential with parameters $(\sqrt{5}\sigma^2, 10\sigma^2)$.

In our setting, $Y_i = \phi_i - \mu_i$ is $\sigma$ sub-Gaussian and $\gamma_i = \mathbf{E}[(\phi_i - \mu_i)^2] \leq \sigma^2$. Applying Proposition 9 to $Y_i^2$, we have

$$\begin{aligned}\mathbf{E}[(\phi_T - \mu_T)^2] &\leq \sigma^2 + 10\sigma^2 I(T;\boldsymbol{Y}^2) + \frac{\sigma^2}{4} \\ &\leq \sigma^2\left(1.25 + 10I(T;\boldsymbol{Y}^2)\right) \\ &\leq \sigma^2\left(1.25 + 10I(T;\boldsymbol{\phi})\right).\end{aligned}$$

where $\boldsymbol{Y}^2 \equiv (Y_1^2, ...Y_m^2)$ and the final step uses the data-processing inequality. □

In the next result, we think of $\boldsymbol{\phi} = (\phi_1, .., \phi_m)$ and $T$ as a collection of estimates and a choice of which one to report made based on *common* data-set $D$, while we think of $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, ..., \tilde{\phi}_m)$ as these same estimates computed on a

fresh replication data-set $\tilde{D}$. The next result bounds the KL-divergence between $\phi_T$ and $\tilde{\phi}_T$, which captures the change in the distribution of the reported result due to performing selection and estimation on a common data-set.

**Proposition 10.** *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*

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

*Proof.*

$$\begin{aligned}
&D\left(\mathbf{P}(\phi_T = \cdot)\,||\,\mathbf{P}(\tilde{\phi}_T = \cdot)\right)\\
&\leq D\left(\mathbf{P}(\phi_T = \cdot, T = \cdot)\,||\,\mathbf{P}(\tilde{\phi}_T = \cdot, T = \cdot)\right)\\
&\leq \sum_{T=1}^{m} \mathbf{P}(T=i) D\left(\mathbf{P}(\phi_T = \cdot|T=i)\,||\,\mathbf{P}(\tilde{\phi}_T = \cdot|T=i)\right)\\
&= \sum_{T=1}^{m} \mathbf{P}(T=i) D\left(\mathbf{P}(\phi_i = \cdot|T=i)\,||\,\mathbf{P}(\phi_i = \cdot)\right)\\
&\leq \sum_{T=1}^{m} \mathbf{P}(T=i) D\left(\mathbf{P}(\boldsymbol{\phi} = \cdot|T=i)\,||\,\mathbf{P}(\boldsymbol{\phi} = \cdot)\right)\\
&= I(T;\boldsymbol{\phi}),
\end{aligned}$$

where both inequalities follow from the data-processing inequality for KL divergence. □

## APPENDIX C
## INFORMATION USAGE ALSO LOWER BOUNDS BIAS

### A. Top-k selection: a lower bound for Corollary 1

Here we show that the bound of Corollary 1 is 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 $\phi_{\tilde{T}}$. Because the average among the $m_0$ group leaders is less than the average among the $\phi_{(1)}, ..., \phi_{(m_0)}$, we have $\mathbf{E}[\phi_{\tilde{T}}] \leq \mathbf{E}[\phi_T]$. Moreover, each group leader converges to $\sigma\sqrt{2\log m/m_0}$ and since the groups are independent, the average $\mathbf{E}[\phi_{\tilde{T}}]$ also converges to $\sigma\sqrt{2\log m/m_0}$.

### B. Maximum of Gaussians: Proof of Proposition 3

Recall the statement of Proposition 3.

**Proposition 3.** *Let $T = \arg\max_{1\leq i\leq m} \phi_i$ where $\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}, I)$. There exist universal numerical constants $c_1 = 1/8$, $c_2 < 2.5$ , $c_3 = 10$, and $c_4 = 1.5$ such that for any $m \in \mathbb{N}$ and $\boldsymbol{\mu} \in \mathbb{R}^m$,*

$$c_1 H(T) - c_2 \leq \mathbf{E}[(\phi_T - \mu_T)^2] \leq c_3 H(T) + c_4.$$

The upper bound above follows by Proposition 2. Here we will focus on establishing the lower bound.

Throughout, we will use the notation $M \triangleq \phi_T = \max_i \phi_i$ and $M_{-i} \triangleq \max_{j\neq i} \phi_j$. We rely on the following facts. The first shows that the maximum of Gaussian random variables is itself a sub-Gaussian random variable. The second establishes a tail bound for normal random variables.

**Fact 2.** *$M \triangleq \max_i \phi_i$ is 1-subgaussian. In particular, $\mathbf{E}[e^{\lambda(M-\mathbf{E}[M])}] \leq e^{\lambda^2/2}$. This implies the variance bound $\mathbf{E}[(M-\mathbf{E}[M])]^2 \leq 1$ and the tail bounds $\mathbf{P}(M \geq E[M] + \lambda) \leq e^{-\lambda^2/2}$. Similarly, $M_{-i}$ is 1-sub–Gaussian for all $i$.*

**Fact 3.** *If $X \sim \mathcal{N}(0,1)$ then for all $x > 0$*

$$\mathbf{P}(X > x) \geq \frac{1}{\sqrt{2\pi}}\left(\frac{x}{x^2+1}\right)e^{-x^2/2}$$

Proposition 3 provides an analogous lower bound. To understand this result, recall that entropy the entropy of $T$ is

$$H(T) = \sum_i \mathbf{P}(T=i)\log(1/\mathbf{P}(T=i)).$$

Consider a setting where $\mathbf{E}[M]$ significantly exceeds $\mu_i$. Then, since $M$ concentrates around $\mathbf{E}[M]$, the probability $i$ is maximal is close to the probability $\phi_i$ exceeds $\mathbf{E}[M]$. By the above fact, one expects that $\log(1/\mathbf{P}(T=i)) \approx \log \mathbf{P}(\phi_i > \mathbf{E}[M]) \approx (\mathbf{E}[M]-\mu_i)^2/2$. This is roughly the intuition behind the following result. Along with our upper bound, this describes a natural family of problems in which $\mathbf{E}[(\phi_T - \mu_T)^2] = \Theta(1 + H(T))$.

*Proof.* We focus on establishing the lower bound, as the upper bound follows from Proposition 2.

By definition, $T = i$ if and only if $M_{-i} \leq \phi_i$. Our proof will separately consider two cases, depending on whether $\mathbf{E}[M_{-i}] \geq \mu_i + 1$. Let $I \equiv \{i : \mathbf{E}[M_{-i}] \geq \mu_i + 1\}$ denote the set of estimates whose mean is at least a full standard deviation below that of $M_{-i}$.

The entropy of $T$ can be decomposed as

$$\begin{aligned}
H(T) =& \sum_{i\notin I} \mathbf{P}(T=i)\log\left(\frac{1}{\mathbf{P}(T=i)}\right)\\
&+ \sum_{i\in I} \mathbf{P}(T=i)\log\left(\frac{1}{\mathbf{P}(T=i)}\right).
\end{aligned}$$

We first upper bound the sum over $i \notin I$. We do this by lower bounding $\mathbf{P}(T=i)$, which yields an upper bound on $\log(1/\mathbf{P}(T=i))$. For any constant $\lambda > 0$, and $i \notin I$, $\mathbf{P}(M_{-i} < \mathbf{E}[M_{-i}] + \lambda) > 1 - e^{-\lambda^2/2}$. Using the fact that $\mathbf{E}[M_{-i}] < \mu_i + 1$, we have for all $\lambda \geq 0$

$$\begin{aligned}
\mathbf{P}(T=i) &= \mathbf{P}(M_{-i} < \phi_i)\\
&\geq \mathbf{P}(M_{-i} < \mathbf{E}[M_{-i}] + \lambda)\cdot\mathbf{P}(\phi_i > \mathbf{E}[M_{-i}] + \lambda)\\
&\geq \mathbf{P}(M_{-i} < \mathbf{E}[M_{-i}] + \lambda)\cdot\mathbf{P}(\phi_i > \mu_i + 1 + \lambda)\\
&\geq \left(1 - e^{-\lambda^2/2}\right)\frac{1}{\sqrt{2\pi}}\left(\frac{1+\lambda}{(1+\lambda)^2+1}\right)e^{-(1+\lambda)^2/2}\\
&\triangleq p(\lambda).
\end{aligned}$$

Therefore

$$\begin{aligned}&\sum_{i\notin I}\mathbf{P}(T=i)\log\left(\frac{1}{\mathbf{P}(T=i)}\right)\\ \leq\;& \mathbf{P}(T\notin I)\max_{i\notin I}\log\left(\frac{1}{\mathbf{P}(T=i)}\right)\\ \leq\;& \log\left(\frac{1}{p(1)}\right)\triangleq c_{-I}.\end{aligned}$$

Direct calculation shows $c_{-I}<5$.

Now we consider the case $i\in I$. To simplify notation, consider the shifted random variables $X\equiv\phi_i-\mu_i\sim\mathcal{N}(0,1)$ and $Y\equiv M_{-i}-\mu_i$. We lower bound $\log(1/\mathbf{P}(T=i))$ by a function of $\mathbf{E}[Y]^2$. We have

$$\begin{aligned}&\mathbf{P}(T=i)\\ =\;&\int_{-\infty}^{\infty}\mathbf{P}(X>x)\mathbf{P}(Y=dx)\\ \geq\;&\int_{1}^{\infty}\mathbf{P}(X>x)\mathbf{P}(Y=dx)\\ =\;&\mathbf{P}(Y\geq 1)\int_{1}^{\infty}\mathbf{P}(X>x)\mathbf{P}(Y=dx|Y\geq 1)\\ \geq\;&\frac{\mathbf{P}(Y\geq 1)}{\sqrt{2\pi}}\int_{1}^{\infty}\left(\frac{x}{x^2+1}\right)e^{-x^2/2}\mathbf{P}(Y=dx|Y\geq 1).\end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}\log\mathbf{P}(T=i)\geq\;&\log(1/\sqrt{2\pi})+\log(\mathbf{P}(Y\geq 1))\\ &+\int_{1}^{\infty}\left(\log\left(\frac{x}{x^2+1}\right)-x^2/2\right)\\ &\times\mathbf{P}(Y=dx|Y\geq 1),\end{aligned}$$

which can be rewritten as

$$\begin{aligned}\log\left(\frac{1}{\mathbf{P}(T=i)}\right)\leq&\log(\sqrt{2\pi})+\log\left(\frac{1}{\mathbf{P}(Y\geq 1)}\right)\\ &+\mathbf{E}\left[\log\left(\frac{Y^2+1}{Y}\right)|Y>1\right]\\ &+\frac{\mathbf{E}[Y^2|Y>1]}{2}.\end{aligned}$$

For $Y\geq 1$, one has $\log((Y^2+1)/Y)\leq\log(1+Y)\leq Y$. Therefore,

$$\begin{aligned}\log\left(\frac{1}{\mathbf{P}(T=i)}\right)\leq&\log(\sqrt{2\pi})+\log\left(\frac{1}{\mathbf{P}(Y\geq 1)}\right)\\ &+1.5\mathbf{E}[Y^2|Y>1].\end{aligned}$$

Now,

$$\begin{aligned}\mathbf{E}[Y^2|Y>1]&\leq\mathbf{E}[Y^2]/\mathbf{P}(Y>1)\\ &=\left(\mathbf{E}[(Y-\mathbf{E}[Y])^2]+\mathbf{E}[Y]^2\right)/\mathbf{P}(Y>1).\end{aligned}$$

Since $Y=M_{-i}-\mu_i$, the variance of $Y$ is bounded by 1. Using as well that $\mathbf{P}(Y>1)\geq 1-1/\sqrt{e}$ gives the bound

$$\begin{aligned}\log\left(\frac{1}{\mathbf{P}(T=i)}\right)\leq&\log(\sqrt{2\pi})+\log\left(\frac{1}{\mathbf{P}(Y\geq 1)}\right)\\ &+\frac{1.5(1+\mathbf{E}[Y]^2)}{\mathbf{P}(Y\geq 1)}\\ <&5+4\mathbf{E}[Y]^2.\end{aligned}$$

Now, plugging in $\mathbf{E}[Y]=\mathbf{E}[M_{-i}]-\mu_i$ and putting everything together, we find

$$\begin{aligned}H(T)&=\sum_i\mathbf{P}(T=i)\log(1/\mathbf{P}(T=i))\\ &\leq c_{-I}+5+4\sum_{i\in I}\mathbf{P}(T=i)(\mathbf{E}[M_{-i}]-\mu_i)^2\\ &\leq c_{-I}+5+4\sum_{i\in I}\mathbf{P}(T=i)(\mathbf{E}[M]-\mu_i)^2\\ &\leq c_{-I}+5+4\|\mathbf{E}[M]-\mu_T\|^2\end{aligned}$$

where $\|X\|\equiv\sqrt{\mathbf{E}[X^2]}$ denotes the $L_2$ norm a random variable $X$ and the second inequality uses that $\mathbf{E}[M]\geq\mathbf{E}[M_{-i}]\geq\mu_i$.

We complete the proof by relating $\|\mathbf{E}[M]-\mu_T\|$ to $\|\phi_T-\mu_T\|$. Recall that $\phi_T$ is 1-sub–Gaussian and $\mathbf{E}[\phi_T]=\mathbf{E}[M]$. Therefore

$$\|\mathbf{E}[M]-\phi_T\|=\mathbf{E}[(\phi_T-\mathbf{E}[\phi_T])^2]\leq 1.$$

Combining this with the triangle inequality shows

$$\|\mathbf{E}[M]-\mu_T\|=\|\mathbf{E}[M]-\phi_T+\phi_T-\mu_T\|\leq 1+\|\phi_T-\mu_T\|.$$

We can then conclude

$$\|\mathbf{E}[M]-\mu_T\|^2\leq(1+\|\phi_T-\mu_T\|)^2\leq 2+2\|\phi_T-\mu_T\|^2$$

where the inequality uses that $\max_{x\in\mathbb{R}}f(x)=0$ for $f(x)\equiv(1+x)^2-2-2x^2$. Together, this shows

$$H(T)\leq c_{-I}+5+8+8\|\phi_T-\mu_T\|^2$$

or

$$\|\phi_T-\mu_T\|^2\geq c_1H(T)-c_2$$

where $c_1=1/8$ and $c_2=(c_{-I}+13)/8<2.5$. □

### C. Threshold Selection with Gaussian Random Variables

In addition to the max-selection policy, we analyze a softer threshold selection policy and prove that the information usage lower bounds bias here as well. Let each $\phi_i$ correspond to a Gaussian of variance 1, and we allow the Gaussians to have different means and be correlated.

Let $M$ be a constant. The threshold-$M$ selection procedure does the following:

1) If at least one $\phi_i$ is larger than $M$, uniformly randomly select one of these $\phi_i$'s to report. For this, we exclude $\phi_{-1}$.
2) Otherwise, always report an arbitrary, fixed $\phi_{-1}$.

In what follows, we will show that for $M$ *sufficiently large*, the entropy $H(T)$ lower bounds the square-loss bias $\mathbf{E}[(Z_T-\mu_T)^2]$, where, recall that $Z_T=\mathbf{E}[\phi_i|T=i]$. Let $N_{-i}=$

$|\{\phi_j \geq M, j \neq i, j \neq -1\}|$. As $M$ increases, $\mathbf{E}[N_{-i}|\phi_i \geq M]$ decreases. We want the threshold to be high enough so that only a few $\phi_i$'s are expected to pass the threshold. Let $\hat{N}(M) = \max_i \mathbf{E}[N_{-i}|\phi_i \geq M]$.

**Theorem 1.** *Suppose*

$$\begin{aligned} & M - \max_i \mu_i \\ \geq & \sqrt{2\log[2\pi\left(1+\mathbf{E}\left[N_{-i}|\phi_i \geq M\right]\right)(M-\max_i \mu_i)]+3}, \end{aligned}$$

*then*

$$\mathbf{E}[(\phi_T - \mu_T)^2] \geq H(T).$$

*Proof.* For $i \neq -1$, define $p_i = \mathbf{P}(T=i)$. Then we have

$$\begin{aligned} p_i &= \mathbf{P}(\phi_i \geq M) \sum_{k=0}^{n-1} \mathbf{P}(N_{-i}=k|\phi_i \geq M)\frac{1}{k+1} \\ &= \mathbf{P}(\phi_i \geq M)\mathbf{E}\left[\frac{1}{1+N_{-i}}|\phi_i \geq M\right]. \end{aligned}$$

Let $p = \sum p_i$ denote the probability that at least one $\phi_i, i \neq -1$, passes the threshold. Note that here and below, when we write $\sum p_i$, we always mean the sum of over $i \neq -1$. We can write the entropy as

$$\begin{aligned} H(T) =& \sum p_i \log \frac{1}{p_i} + (1-p)\log\frac{1}{1-p} \\ =& \sum p_i \log \frac{1}{\mathbf{P}(\phi_i \geq M)} \\ & + \sum p_i \log\left(1/\mathbf{E}\left[\frac{1}{1+N_{-i}}|\phi_i \geq M\right]\right) \\ & + (1-p)\log\frac{1}{1-p} \\ \leq& \sum p_i \log \frac{1}{\mathbf{P}(\phi_i \geq M)} \\ & + \sum p_i \log\left(1+\mathbf{E}\left[N_{-i}|\phi_i \geq M\right]\right) \\ & + (1-p)\log\frac{1}{1-p} \\ \leq& \sum p_i \log \frac{1}{\mathbf{P}(\phi_i \geq M)} \\ & + \sum p_i \log\left(1+\mathbf{E}\left[N_{-i}|\phi_i \geq M\right]\right) + p. \end{aligned}$$

We can rewrite the inequality as

$$\begin{aligned} \sum p_i \log \frac{1}{\mathbf{P}(\phi_i \geq M)} \geq & H(T) \\ & - \sum p_i \log\left(1+\mathbf{E}\left[N_{-i}|\phi_i \geq M\right]\right) \\ & - p. \end{aligned}$$

Since $\phi_i \sim \mathcal{N}(\mu_i, 1)$ and $M > \mu_i$, we have the bounds

$$\begin{aligned} \frac{(M-\mu_i)^2}{2} \geq \log\frac{1}{\mathbf{P}(\phi_i \geq M)} - \log(M-\mu_i) - \frac{1}{2}\log(2\pi) \\ - \frac{1}{(M-\mu_i)^2}. \end{aligned}$$

After some algebra we have

$$\begin{aligned} & \mathbf{E}[(\phi_T - \mu_T)^2] \\ \geq & \sum p_i (M-\mu_i)^2 \\ \geq & \sum p_i \left[\frac{(M-\mu_i)^2}{2} - \log\left(1+\mathbf{E}\left[N_{-i}|\phi_i \geq M\right]\right)\right. \\ & \left. - \frac{\log 2\pi}{2} - \log(M-\mu_i) - \frac{1}{(M-\mu_i)^2} - 1\right] \\ & + H(T) \\ \geq & H(T) \end{aligned}$$

where the second inequality used the above inequalities for $\frac{(M-\mu_i)^2}{2}$ and $\sum p_i \log \frac{1}{\mathbf{P}(\phi_i \geq M)}$; and the third inequality used the condition that $M - \max_i \mu_i$ exceeds $\sqrt{2\log[2\pi\left(1+\mathbf{E}\left[N_{-i}|\phi_i \geq M\right]\right)(M-\max_i \mu_i)]+3}$. □

As $M$ increases, unless the $\phi_i$'s are very highly correlated, $\mathbf{E}[N_{-i}|\phi_i \geq M])$ decreases and $H(T)$ dominates in the inequality. This shows that $H(T)$ is a natural lower bound on $\mathbf{E}(Z_T^2)$ and hence $\sqrt{H(T)}$ lower bounds bias. Actually we can improve this lower bound by considering $I(T|\Phi) = H(T) - H(T|\Phi)$ using the fact that

$$\begin{aligned} H(T|\Phi) &= \sum_{N=1} \mathbf{P}(N=i) H(T|N=i) \\ &= \sum_{N=1} \mathbf{P}(N=i)\log i \\ &= \mathbf{E}[\log N | N \geq 1] \end{aligned}$$

where $N = |\{\phi_i, \phi_i \geq M\}|$. Assuming that $\phi_i$'s are independent, we need to control the gap between $\mathbf{E}[Z_T^2]$ and $I(T,\Phi)$, we need to upper bound $p\log(1+\mathbf{E}[N]) - \mathbf{E}[\log N|N \geq 1]$.

### *D. Threshold Selection with Exponential Random Variables*

We can prove the analogous lower bound for the threshold policy with exponential random variables. Let $\phi_i = \lambda_i + \exp(1)$ be the shifted exponential random variable. So for $x \geq \lambda_i$, $\mathbf{P}(\phi_i = x) = e^{-(x-\lambda_i)}$ and $\mathbf{P}(\phi_i = x) = 0$ for $x < \lambda_i$. Different $\phi_i$'s can have different $\lambda_i$ and we allow them to be correlated. The mean of $\phi_i$ is $\mu_i = \lambda_i + 1$. As before, let $\hat{N}(M) = \max_i \mathbf{E}[N_{-i}|\phi_i \geq M]$.

**Theorem 2.** *Suppose* $M - \max \lambda_i \geq 4 + 2\log(1+\hat{N}(M))$,

$$\mathbf{E}[\phi_T - \mu_T] \geq H(T)/2.$$

*Proof.* The proof follows the same structure as before. Since $\mathbf{P}(\phi_i > M) = e^{-(M-\lambda_i)}$, we have $\log 1/\mathbf{P}(\phi_i > M) = M - \lambda_i$ and

$$\begin{aligned} H(T) = & \sum_{i \neq -1} p_i (M-\lambda_i) \\ & + \sum_{i \neq -1} p_i \log\left(1/\mathbf{E}\left[\frac{1}{1+N_{-i}}|\phi_i \geq M\right]\right) \\ & + (1-p)\log\frac{1}{1-p}. \end{aligned}$$

On the other hand,

$$\begin{aligned}
& \mathbf{E}[\phi_T - \mu_T] \\
\geq & \sum_{i \neq -1} p_i (M - \mu_i) \\
= & \sum_{i \neq -1} p_i (M - \lambda_i) - p \\
\geq & H(T) - \sum_{i \neq -1} \log \left( 1/\mathbf{E} \left[ \frac{1}{1+N_{-i}} | \phi_i \geq M \right] \right) - 2p \\
\geq & H(T) - \sum_{i \neq -1} \log(1 + \mathbf{E}[N_{-i} | \phi_i \geq M]) - 2p \\
\geq & H(T)/2.
\end{aligned}$$

□

## APPENDIX D
## INFORMATION USAGE AND CLASSIFICATION OVERFITTING: PROOF OF PROP. 5

*Proof.* The empirical $\hat{L}(f)$ and expected $L(f)$ loss of a classifier $f \in \mathcal{F}$ on the training examples $\mathbf{x} \equiv (x_1, ..., x_n)$ depend only on the predictions $f(\mathbf{x}) \equiv (f(x_1), ..., f(x_n))$ it makes on these examples. Let $\mathcal{F}_{\mathbf{x}} = \{f(\mathbf{x}) : f \in \mathcal{F}\}$ and note that $m = |\mathcal{F}_{\mathbf{x}}| \leq 2^n$ is finite. Let $f_1, ..., f_m$ be functions that make different classifications at $\mathbf{x}$, so $\cup_1^m \{f(\mathbf{x})\} = \mathcal{F}_{\mathbf{x}}$.

Now, the overfitting problem studied in Prop. 5 can be cast in the same framework as the rest of the paper. For each $i \in \{1, ..., m\}$, set $\phi_i = \hat{L}(f_i)$ and $\mu_i = L(f_i)$ to be the training error and expected error of classifier $f_i$. Let $T \in \{1, .., m\}$ be the random index satisfying $\hat{f}(\mathbf{x}) = f_T(\mathbf{x})$. Then, our result follows by bounding $|\mathbf{E}[\phi_T - \mu_T]|$.

If $X \sim \mathrm{Bern}(p)$ is a Bernoulli random variable with parameter $p$, then $X - p$ is sub-Gaussian with parameter less than 1/4 [36]. Similarly, if $X_1, ..X_n$ are Bernoulli random variables with respective parameters $p_1, ..., p_n$, then $n^{-1} \sum_{i=1}^n (X_i - p_i)$ is sub-Gaussian with parameter not exceeding $1/4\sqrt{n}$. This immediately implies $\phi_i - \mu_i$ is $\sigma$–sub-Gaussian with $\sigma \leq 1/2n$, so applying Prop. 1 implies

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

Using the information-processing inequality, and the definition of $T$, we have

$$I(T;\boldsymbol{\phi}) \leq I(T;\mathbf{Y}) = I(\hat{f}(\mathbf{x}), \mathbf{Y}),$$

which completes the proof of the first claim.

The second claim uses a standard link between VC-dimension and the size of $\mathcal{F}_{\mathbf{x}}$. Set $S_{\mathcal{F}}(n) = \max\{|\mathcal{F}_{\mathbf{x}}| : \mathbf{x} \in \mathcal{X}^n\}$ to be the maximum number of ways $n$ points can be classified by the function class. This is often called the growth function of $\mathcal{F}$. We have immediately that $I(T;\boldsymbol{\phi}) \leq H(T) \leq \log S_{\mathcal{F}}(n)$. Sauer's lemma (see Lemma 1 of [37]) shows that

$$S_{\mathcal{F}}(n) \leq \begin{cases} 2^n & \text{if } n < d \\ \left(\frac{en}{d}\right)^d & \text{if } n \geq d \end{cases}$$

where $d < \infty$ is the VC-dimension of $\mathcal{F}$. This implies $\log S_{\mathcal{F}}(n) \leq d \log_+ \left(\frac{en}{d}\right)$. □

## APPENDIX E
## BIAS CONTROL VIA FDR CONTROL: PROOF OF PROP. 6

*Proof.* Let $\chi = \mathbf{1}_{\{T \in \mathcal{H}_0\}}$ denote an indicator for a false discovery. Since this is a deterministic function of $T$,

$$I(T;\phi) = I(T;(\chi,\boldsymbol{\phi})) = I(T;\chi) + I(T;\boldsymbol{\phi} \mid \chi) \leq H(\chi) + I(T;\boldsymbol{\phi} \mid \chi)$$

Now, using the distribution of $\chi$ we have

$$\begin{aligned}
I(T;\boldsymbol{\phi} \mid \chi) \leq & H(\chi) + \mathbf{P}(\chi = 0) I(T;\boldsymbol{\phi} \mid \chi = 0) \\
& + \mathbf{P}(\chi = 1) I(T;\boldsymbol{\phi} \mid \chi = 1) \\
= & h(\mathbf{P}(T \in \mathcal{H}_1)) + \mathbf{P}(T \in \mathcal{H}_1) I(T;\boldsymbol{\phi} \mid T \in \mathcal{H}_1) \\
& + \mathbf{P}(T \in \mathcal{H}_0) I(T;\boldsymbol{\phi} \mid T \in \mathcal{H}_0) \\
\leq & h(\mathbf{P}(T \in \mathcal{H}_1)) + \mathbf{P}(T \in \mathcal{H}_1) I(T; D \mid T \in \mathcal{H}_1) \\
& + \mathbf{P}(T \in \mathcal{H}_0) I(T; D \mid T \in \mathcal{H}_0)
\end{aligned}$$

where we have applied chain rule and the data-processing inequality. Now, we have

$$\begin{aligned}
I(T; D \mid T \in \mathcal{H}_1) = & H(T \mid T \in \mathcal{H}_1) - H(T | T \in \mathcal{H}_1, D) \\
\leq & \log(\#\mathcal{H}_1) \\
& - \mathbf{E}[\log(\#(S_1 \cap \mathcal{H}_1)) \mid T \in \mathcal{H}_1] \\
= & \log(\#\mathcal{H}_1) \\
& - \mathbf{E}[\log\left((1-\hat{\beta}) \cdot (\#\mathcal{H}_1)\right) \mid T \in \mathcal{H}_1] \\
= & -\mathbf{E}\left[\log\left(1-\hat{\beta}\right) \mid T \in \mathcal{H}_1\right] \\
= & -\log(1-\beta) + \mathbf{E}\left[\log\left(\frac{1-\beta}{1-\hat{\beta}}\right) \mid T \in \mathcal{H}_1\right].
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbf{P}(T \in \mathcal{H}_1) I(T; D \mid T \in \mathcal{H}_1) \leq & -\mathbf{P}(T \in \mathcal{H}_1) \log(1-\beta) \\
& + \mathbf{E}\left[\log\left(\frac{1-\beta}{1-\hat{\beta}}\right) \mathbf{1}_{\{T \in \mathcal{H}_1\}}\right] \\
\leq & -\mathbf{P}(T \in \mathcal{H}_1) \log(1-\beta) \\
& + \mathbf{E}\left[\log_+\left(\frac{1-\beta}{1-\hat{\beta}}\right)\right].
\end{aligned}$$

Essentially the same calculation shows,

$$\mathbf{P}(T \in \mathcal{H}_0) I(T; X \mid T \in \mathcal{H}_0) \leq -\mathbf{P}(T \in \mathcal{H}_0) \log(\alpha) + \mathbf{E}\left[\log_+\left(\frac{\alpha}{\hat{\alpha}}\right)\right].$$

Plugging in $\mathbf{P}(T \in \mathcal{H}_0) = \mathrm{FDR}$ concludes the proof. □

## APPENDIX F
## ADDITIONAL APPLICATIONS OF INFORMATION USAGE

When a data analyst selects hypothesis tests to perform after data exploration, they may compute extremely small p-values even if there is no signal in the data, and all null hypotheses hold. In this section we apply our information-usage framework to quantify how severely the analyst must explore the data to produce these small p-values. We also give an illustration of mutual information bound in controlling the value of information in decision-making under uncertainty.

### A. The Probability of Small p–values

Let $\phi_i$ be the observed $p$-value of the $i$th hypothesis and suppose the analyst has 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 $T$.

In fact, the bound provided by the following proposition is stronger. Instead of depending on $I(T; \boldsymbol{\phi})$, it depends on the mutual information between $T$ and a more compressed random variable $\mathbf{Z}_\epsilon$. Here $Z_{\epsilon,i} \equiv \mathbf{1}(\phi_i < \epsilon)$ and the term $I(T; \mathbf{Z}_\epsilon) \leq I(T; \boldsymbol{\phi})$ is a measure of the dependence of the selection rule on the realization of extremely small p-values.

**Proposition 11.** *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*

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

*Proof of Proposition 11.* Since $\phi_i \sim \text{Uniform}(0, 1)$, $Z_{\epsilon,i} = \mathbf{1}(\phi_i < \epsilon)$ is a Bernoulli random variable with parameter $\epsilon$ and $\mathbf{E}[Z_i] = \epsilon$. We use the fact [36] that a probability $p$ Bernoulli random variable is sub-Gaussian with parameter

$$\sigma = \sqrt{\frac{1 - 2p}{2\log((1 - p)/p)}} \leq \sqrt{\frac{1}{2\log(1/2p)}}.$$

Combining this with Proposition 1, we have the desired result

$$\mathbf{E}[Z_T] - \mathbf{E}[\mu_T] = \mathbf{P}(p_T < \epsilon) - \epsilon \leq \sqrt{\frac{I(T; \mathbf{Z}_\epsilon)}{\log(1/2\epsilon)}}.$$

□

To interpret this result, suppose the selection procedure $T$ reports the minimal $p$-value and $\epsilon = 2^{-k}$. If we test $2^k$ independent hypotheses, then standard multiple hypotheses testing theory tells us that there is a non-neglible probability that $p_T$ is less than $\epsilon$. This shows up in the bound of Proposition 11 since $\sqrt{\frac{I(T;\mathbf{Z}_\epsilon)}{\log(1/2\epsilon)}} \approx 1$. However, when there is correlation among the hypotheses, $I(T; \mathbf{Z}_\epsilon)$ can be significantly less than $2^k$, and our bound quantifies the risk of false discovery in this more nuanced setting.

### B. Regret Analysis and the 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 12.** *If for each for each $x \in \mathcal{X}$, $f_\theta(x) - \mu(x)$ is $\sigma$ sub-Gaussian, then*

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

*Proof.* Note that

$$\max_{x \in \mathcal{X}} \mu(x) \geq \mathbf{E}[\mu(X^*)]$$

and

$$\mathbf{E}[\max_{x \in \mathcal{X}} f_\theta(x)] = \mathbf{E}[f_\theta(X^*)]$$

Therefore,

$$\begin{aligned} \mathbf{E}[\max_{x \in \mathcal{X}} f_\theta(x)] - \max_{x \in \mathcal{X}} \mu(x) &\leq \mathbf{E}[f_\theta(X^*)] - E[\mu(X^*)] \\ &\leq \sigma\sqrt{2I(X^*; \theta)} \\ &= \sigma\sqrt{2H(X^*)} \end{aligned}$$

□

## APPENDIX G
## ADITIONAL EXPERIMENTAL DETAILS

Here we provide additional details for the LARS bias experiments of Figure 2. We consider random design matrix $X \in \mathbb{R}^{100 \times 1000}$ whose entries are i.i.d. samples from $\mathcal{N}(0, 1)$. The rows of $X$ are then normalized to have unit variance. The effects are represented by the vector $\beta \in \mathbb{R}^{1000}$. The first 20 entries of $\beta$ are set to a constant $s$—corresponding to the signals—and rest of the entries are all set to be 0. By increasing $s$, we increase the signal-to-noise in the data. The

*low*, *medium* and *high* signal settings corresponds to setting $s = 0.04, 0.06$ and $0.08$, respectively. Finally the outcomes are given by $y = X \cdot \beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, I_{100}/10)$ is the noise. We consider the full selection path of LARS on $X$ and $y$. Let the index $T_i$ denote the $i$th feature to enter the subset selected by LARS.

In this experiment, for simplicity, we quantify the bias on the *univariate* regression coefficients. More concretely, suppose we have the true values $y^* = X \cdot \beta$. Then we can use least squares between $y^*$ and the $j$th column of $X$ to determine the true univariate coefficient $\beta_j^*$ of the feature $j$. From the noisy observations $y$, we can similarly compute the noisy univariate coefficient $\hat{\beta}_j$. We quantify the bias $\hat{\beta}_{T_i} - \beta^*_{T_i}$, for $i = 1, 2, ....$ This bias quantifies how much LARS overfit to the noise in the data.

## APPENDIX H
## COMPLETE ANALYSIS OF THE MULTI-STEP DATA ANALYSIS MODEL

*Proof of Lemma 1.* Since, conditional on $H_k$, $T_{k+1}$ is independent of $\boldsymbol{\phi}$, the data-processing inequality for mutual information implies,

$$I(T_{k+1}; \boldsymbol{\phi}) \leq I(H_k; \boldsymbol{\phi}).$$

Now we have,

$$I(H_k; \boldsymbol{\phi}) = \sum_{i=1}^{k} I\left((T_i, Y_{T_i}); \boldsymbol{\phi} | H_{i-1}\right).$$

We complete the proof by simplifying the expression for $I\left((T_i, Y_{T_i}); \boldsymbol{\phi} | H_{i-1}\right)$. Let $\boldsymbol{\phi}_{(-i)} = (\phi_j : j \neq i)$. Then,

$$\begin{aligned} I\left((T_i, Y_{T_i}); \boldsymbol{\phi} | H_{i-1}\right) &= I\left(T_i; \boldsymbol{\phi} | H_{i-1}\right) \\ &\quad + I\left(Y_{T_i}; \boldsymbol{\phi} | H_{i-1}, T_i\right) \\ &= I\left(Y_{T_i}; \boldsymbol{\phi} | H_{i-1}, T_i\right) \\ &= I(Y_{T_i}; \phi_{T_i} | H_{i-1}, T_i) \\ &\quad + I(Y_{T_i}; \phi_{(-T_i)} | H_{i-1}, T_i, \boldsymbol{\phi}_{T_i}) \\ &= I(Y_{T_i}; \phi_{T_i} | H_{i-1}, T_i), \end{aligned}$$

where the final equality follows because, conditioned on $\phi_{T_i}$, $Y_{T_i}$ is independent of $\boldsymbol{\phi}_{(-T_i)}$. □

*Proof of Lemma 2.*

$$\begin{aligned} I(X; Y) &= -\frac{1}{2} \log\left(1 - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right) \\ &= -\frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \\ &= \frac{1}{2} \log\left(1 + \frac{\sigma_1^2}{\sigma_2^2}\right). \end{aligned}$$

□

**Lemma 3.** *Let $X$ be a real value random variable with variance $\sigma_X^2 = (X - \mathbf{E}[X])^2]$ and $W \sim \mathcal{N}(0, \sigma_W^2)$ be a normal random variable that is independent of $X$. Then*

$$I(X; X + W) \leq \frac{\sigma_X^2}{\sigma_W^2}$$

*Proof.* Let $p_X(x)$ denote the density of $X$ with respect to some base measure $\nu$ over $\mathcal{X}$. Then we have

$$\begin{aligned} & I(X; X + W) \\ = & \int_{\mathcal{X}} D(\mathbf{P}(x + W = \cdot) \,||\, \mathbf{P}(X + W = \cdot)) p_X(x) d\nu(x) \\ \overset{(a)}{\leq} & \int_{\mathcal{X}} \left[ \int_{\mathcal{X}} D(\mathbf{P}(x_1 + W = \cdot) \,||\, \mathbf{P}(x_2 + W = \cdot)) p_X(x_2) d\nu(x_2) \right] \\ & \times p_X(x_1) d\nu(x_1) \\ \overset{(b)}{=} & \int_{\mathcal{X}} \int_{\mathcal{X}} \frac{(x_1 - x_2)^2}{2\sigma_W^2} p_X(x_1) p_X(x_2) d\nu(x_1) d\nu(x_2) \\ \overset{(c)}{=} & \frac{\sigma_X^2}{\sigma_W^2}. \end{aligned}$$

□

*Here inequality (a) uses the convexity of KL divergence, (b) follows from the formula for the KL divergence between univariate normal distributions $\mathcal{N}(x_1, \sigma^2)$ and $\mathcal{N}(x_2, \sigma^2)$, and (c) uses that if $X_1$ and $X_2$ are iid random variables with mean $\mu$, then*

$$\mathbf{E}[(X_1 - X_2)^2] = \mathbf{E}[(X_1 - \mu + \mu - X_2)^2] = 2\mathbf{E}[(X_1 - \mu)^2].$$

We prove a more general statement of Proposition 7.

**Proposition 13.** *Suppose $\phi_i \sim \mathcal{N}(\mu_i, \frac{\sigma^2}{n})$ and $(\phi_1, ..., \phi_k)$ is jointly Gaussian for any $k$. If for the $j$th query, $Y_{T_j} = \phi_{T_j} + W_j$ where $W_j \sim \mathcal{N}(0, \frac{\omega_j^2}{n})$ and $(W_1, W_2, ...)$ is independent of $\boldsymbol{\phi}$, then*

$$\mathbf{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] \leq \frac{\sigma}{\sqrt{n}} + c_1 \left( \frac{\omega_{k+1}}{\sqrt{n}} + \sigma^2 \sqrt{\frac{\sum_{j=1}^{k} w_j^{-2}}{n}} \right).$$

*If $\omega_j = \sigma j^{1/4}$ for each $j \in \mathbb{N}$, then for every $k \in \mathbb{N}$*

$$\mathbf{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] \leq c_2 \left( \frac{\sigma k^{1/4}}{n^{1/2}} \right)$$

*where $c_1$ and $c_2$ denote universal constants that are independent of $\sigma, \omega, k$, and $n$.*

*Proof of Proposition 13.*

$$\begin{aligned} \mathbf{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] &\leq \mathbf{E}[|Y_{T_{k+1}} - \phi_{T_{k+1}}|] + \mathbf{E}[|\phi_{T_{k+1}} - \mu_{T_{k+1}}|] \\ &\leq \sqrt{\frac{2\omega_{k+1}}{\pi n}} + \mathbf{E}[|\phi_{T_{k+1}} - \mu_{T_{k+1}}|] \\ &\leq \sqrt{\frac{2\omega_{k+1}}{\pi n}} + \frac{\sigma}{\sqrt{n}} + c \cdot \sigma \sqrt{\frac{2I(T_{k+1}; \boldsymbol{\phi})}{n}} \end{aligned}$$

where $c$ is a universal numerical constant. The second inequality uses the expected value of the half-normal distribution, and the third inequality follows from Proposition 2.

The desired result follows by bounding the mutual information term. Applying Lemma 1, we have

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

where $H_k = (T_1, Y_{T_1}, T_2, Y_{T_2}, ..., T_k, Y_{T_k})$ denotes the history of interaction up to time $k$. Because the $\phi_i$'s are jointly Gaussian, and observation noise is Gaussian, the posterior $\mathbf{P}(\phi_j = \cdot | H_{i-1})$ is Gaussian with conditional variance less than $\sigma^2/n$. Moreover, conditional on $H_{i=1}$, $T_i$ is independent of $(\phi_1, \phi_2...)$ and $(Y_1, Y_2, ....)$, so $\phi_{T_i}|H_{i-1}, T_i$ is normally distributed with variance less than $\sigma^2/n$.

Lemma 2 implies

$$I(Y_{T_i}; \phi_{T_i} | H_{i-1}, T_i) \leq \frac{\sigma^2/n}{2\omega_i^2/n} = \frac{\sigma^2}{2\omega_i^2}$$

and therefore

$$I(T_{k+1}; \boldsymbol{\phi}) \leq \left(\frac{\sigma^2}{2}\right) \sum_{i=1}^{k} \omega_i^{-2}.$$

Plugging this into the earlier bound implies

$$\mathbf{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] \leq \sqrt{\frac{2\omega_{k+1}}{\pi n}} + \frac{\sigma}{\sqrt{n}} + c\sigma^2 \sqrt{\frac{\sum_{i=1}^{k} \omega_i^{-2}}{n}},$$

which is the desired result. □

**Example 1** (Adaptively fitting a linear model [28])**.** *A data-analyst collects $n$ samples of $\theta_1, ...\theta_n \overset{iid}{\sim} \mathcal{D}$ of $k$ dimensional vectors drawn from an unknown distribution and $\hat{\theta}$ is the average of the $\theta_i$'s. She would like to find a unit vector $x$ that is highly correlated with this distribution, in the sense that $\mathbf{E}_{\theta \sim \mathcal{D}}[x^T \hat{\theta}]$ is large. To do this, she looks to maximize $x^T \hat{\theta}$.*

*Suppose $\mathcal{D} = \mathcal{N}(0, \sigma^2 I)$, so $\mathbf{E}_{\theta \sim \mathcal{D}}[x^T \theta] = 0$ for all $x$. Nevertheless, the analyst can still find a vector with a large inner product with $\hat{\theta}$. Imagine she collects $k$ measurements of $\hat{\theta}$, allowing her to completely uncover the vector, and then chooses $X = \hat{\theta}/\|\hat{\theta}\|$ Then, since $\hat{\theta} \sim \mathcal{N}(0, \frac{\sigma^2}{n} I)$,*

$$\mathbf{E}[X^T \hat{\theta}] = \mathbf{E}[\|\hat{\theta}\| = \Theta\left(\sigma \sqrt{\frac{k}{n}}\right).$$

## APPENDIX I
## MUTUAL-INFORMATION VS MAX-INFORMATION

Recent work has proposed max-information [19], and its generalization, approximate max-information, as a metric to control the error of a worst-case, adversarial, data analyst. This notion was motivated by techniques from differential privacy, which shows that a differentially private mechanism have low approximate max-information, and hence has low error even when the analyst is adversarial.

To understand the relationship between mutual–information and max–information, we revisit the rank selection example from Section V. While max–information provides a powerful tool for analyzing the behavior of a worst-case adaptive protocol, this example shows it can exhibit counter-intuitive behavior when analyzing specific selection procedures.

We assume

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

where $\mu \geq 0$. The analyst selects $T = \arg\max_i \phi_i$. As discussed in Section V, bias decreases as the signal strength $\mu$ increases, and this follows transparently from our information theoretic bound. Indeed, as $\mu$ grows $T$ concentrates on $I^*$, and

$$I(T; \boldsymbol{\phi}) = H(T) = \sum_{i=1}^{m} \mathbf{P}(T = i) \log\left(\frac{1}{\mathbf{P}(T = i)}\right)$$

decreases. This scaling is intuitive. As $T$ concentrates on $I^*$ the selection protocol becomes less and less adaptive, and hence we expect both the selection bias as well as the bias bound which depends on $I(T; \boldsymbol{\phi})$ to decrease.

In contrast max-information has the opposite scaling in this setting: it increases as the signal $\mu$ increases and bias decreases. In fact,

$$\begin{aligned} I_\infty(T; \boldsymbol{\phi}) &= \max_{i, \boldsymbol{y}} \log\left(\frac{\mathbf{P}(\boldsymbol{\phi} = \boldsymbol{y}, T = i))}{\mathbf{P}(\boldsymbol{\phi} = \boldsymbol{y})\mathbf{P}(T = i)}\right) \\ &= \max_{i, \boldsymbol{y}} \log\left(\frac{\mathbf{P}(T = i | \boldsymbol{\phi} = \boldsymbol{y}))}{\mathbf{P}(T = i)}\right) \\ &= \max_{i} \log\left(\frac{1}{\mathbf{P}(T = i)}\right), \end{aligned}$$

where the maximum is over $\boldsymbol{y} \in \mathbb{R}^m$ and is attained for any $\boldsymbol{y}$ with $i = \arg\max_j y_j$. By symmetry, $I_\infty(T; \boldsymbol{\phi}) = \log\left(\frac{m-1}{\mathbf{P}(T \neq I^*)}\right)$, which *increases* as the the probability of selecting $I^*$ grows. Therefore, max–information is minimized when the data analyst inappropriately uses rank-selection even though there is no signal in the data ($\mu = 0$). As $\mu$ increases, so the data-analyst detects $I^*$ with probability tending to 1, max-information increases toward infinity.

The related notion of approximate max-information can exhibit similar counter-intuitive behavior. Following [18], the approximate max-information at level $\beta$ is defined to be

$$I_\infty^\beta(T; \boldsymbol{\phi}) := \max_{\substack{\mathcal{O} \subset [m] \times \mathbb{R}^m \\ \mathbf{P}((T, \boldsymbol{\phi}) \in \mathcal{O}) \geq \beta}} \log\left(\frac{\mathbf{P}((T; \boldsymbol{\phi}) \in \mathcal{O}) - \beta}{\mathbf{P}((T; \tilde{\boldsymbol{\phi}}) \in \mathcal{O})}\right).$$

**Lemma 4.** *If $T = f(\boldsymbol{\phi})$ is a deterministic function of $\boldsymbol{\phi}$, then*

$$I_\infty^\beta(T; \boldsymbol{\phi}) \geq \max_{\substack{i \leq m \\ \mathbf{P}(T = i) \geq 2\beta}} \log\left(\frac{1}{\mathbf{P}(T = i)}\right) - \log(2)$$

*for any $i \in \{1, ...m\}$ with $\mathbf{P}(T = i) \geq 2\beta$.*

*Proof.* Let $\tilde{\boldsymbol{\phi}}$ denote a random variable drawn from the marginal distribution of $\boldsymbol{\phi}$, but drawn independently of $T$. Define $\Phi_i = \{x \in \mathbb{R}^m : f(x) = i\}$ to be the decision region corresponding to element $i$. Then

$$\mathbf{P}(T = i, \boldsymbol{\phi} \in \Phi_i) = \mathbf{P}(T = i)\mathbf{P}(\boldsymbol{\phi} \in \Phi_i | T = i) = \mathbf{P}(T = i)$$

whereas

$$\mathbf{P}(T = i, \tilde{\boldsymbol{\phi}} \in \Phi_i) = \mathbf{P}(T = i)\mathbf{P}(\tilde{\boldsymbol{\phi}} \in \Phi_i) = \mathbf{P}(T = i)^2.$$

If $\mathbf{P}(T=i) \geq 2\beta$, then $\mathcal{O} = \{(i,x) : x \in \Phi_i\}$ is feasible, and therefore

$$\begin{aligned} I_\infty^\beta(T;\boldsymbol{\phi}) &\geq \log\left(\frac{\mathbf{P}(T=i, \boldsymbol{\phi} \in \Phi_i) - \beta}{\mathbf{P}(T=i, \tilde{\boldsymbol{\phi}} \in \Phi_i)}\right) \\ &= \log\left(\frac{\mathbf{P}(T=i) - \beta}{\mathbf{P}(T=i)^2}\right) \\ &\geq \log\left(\frac{\frac{1}{2}\mathbf{P}(T=i)}{\mathbf{P}(T=i)^2}\right) \\ &= \log\left(\frac{1}{\mathbf{P}(T=i)}\right) - \log(2). \end{aligned}$$

□

When there is signal in the data, $\mathbf{P}(T=i)$ is small for those $\phi_i$ that do not have signal (i.e. a true null). When $\beta$ is sufficiently small so that $\mathbf{P}(T=i) \geq 2\beta$, the above lemma shows that $I_\infty^\beta(T;\boldsymbol{\phi})$ can be large, and can increase as $\mathbf{P}(T=i)$ deviates farther from the uniform distribution.

## ACKNOWLEDGMENT

The authors would like to thank John Duchi, Cynthia Dwork, Vitaly Feldman, Aaron Roth, Adam Smith, Thomas Steinke, David Tse and Tsachy Weissman for feedback. J.Z. is supported by NSF AF 1763191 and grants from the Chan-Zuckerberg Initiative.

## REFERENCES

[1] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological science*, p. 0956797611417632, 2011.

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

[3] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001.

[4] A. Belloni, V. Chernozhukov, and C. Hansen, "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, vol. 81, no. 287, pp. 608–650, 2014.

[5] S. Van de Geer, P. Buhlmann, Y. Ritov, R. Dezeure *et al.*, "On asymptotically optimal confidence regions and tests for high-dimensional models," *The Annals of Statistics*, vol. 42, no. 3, pp. 1166–1202, 2014.

[6] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2869–2909, 2014.

[7] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani, "A significance test for the lasso," *Annals of Statistics*, vol. 42, no. 2, p. 413, 2014.

[8] W. Fithian, D. Sun, and J. Taylor, "Optimal inference after model selection," *arXiv preprint arXiv:1410.2597*, 2014.

[9] J. Taylor and R. J. Tibshirani, "Statistical learning and selective inference," *Proceedings of the National Academy of Sciences*, vol. 112, no. 25, pp. 7629–7634, 2015.

[10] J. Taylor, R. Lockhart, R. J. Tibshirani, and R. Tibshirani, "Exact post-selection inference for forward stepwise and least angle regression," *arXiv preprint arXiv:1401.3889*, 2014.

[11] J. D. Lee, D. L. Sun, Y. Sun, J. E. Taylor *et al.*, "Exact post-selection inference, with application to the lasso," *The Annals of Statistics*, vol. 44, no. 3, pp. 907–927, 2016.

[12] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, vol. 2, no. Mar, pp. 499–526, 2002.

[13] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi, "General conditions for predictivity in learning theory," *Nature*, vol. 428, no. 6981, pp. 419–422, 2004.

[14] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2635–2670, 2010.

[15] D. McAllester, "A pac-Bayesian tutorial with a dropout bound," *arXiv preprint arXiv:1307.2118*, 2013.

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

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

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

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

[20] R. Bourgon, R. Gentleman, and W. Huber, "Independent filtering increases detection power for high-throughput experiments," *Proceedings of the National Academy of Sciences*, vol. 107, no. 21, pp. 9546–9551, 2010.

[21] S. Wu, A. Joseph, A. S. Hammonds, S. E. Celniker, B. Yu, and E. Frise, "Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks," *Proceedings of the National Academy of Sciences*, p. 201521171, 2016.

[22] J. Zou, C. Lippert, D. Heckerman, M. Aryee, and J. Listgarten, "Epigenome-wide association studies without the need for cell-type composition," *Nature Methods*, pp. 309–11, 2014.

[23] 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.

[24] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," *arXiv*

*preprint arXiv:1304.6133*, 2013.

[25] S. Kamath and C. Nair, "The strong data processing constant for sums of iid random variables," in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 2550–2552.

[26] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 35–55, 2016.

[27] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[28] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Preserving statistical validity in adaptive data analysis," in *STOC 2015*. ACM, 2014.

[29] C. Genovese and L. Wasserman, "Operating characteristics and extensions of the false discovery rate procedure," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 499–517, 2002.

[30] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 1069–1109, 2011.

[31] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge University Press, 2006.

[32] R. MacCoun and S. Perlmutter, "Blind analysis: Hide results to seek the truth." *Nature*, vol. 526, no. 7572, pp. 187–189, 2015.

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

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

[35] M. Wainwright, "Basic tail and concentration bounds," 2015.

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

[37] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," in *Advanced lectures on machine learning*. Springer, 2004, pp. 169–207.

**Daniel Russo** Daniel Russo received his PhD from Stanford University in 2015. He was a postdoc at Microsoft Research from 2015 to 2016 and an assistant professor at Northwestern's Kellogg School of Management from 2016 to 2017. He is currently an assistant professor in the Decision, Risk, and Operations Division at Columbia Business School.

**James Zou** James Zou received his Ph.D. in applied mathematics from Harvard University in 2014. He was a postdoc at Microsoft Research from 2014 to 2016 and is currently an assistant professor at Stanford University.