The Annals of Statistics
2023, Vol. 51, No. 4, 1666–1691
https://doi.org/10.1214/23-AOS2303

POST-SELECTION INFERENCE VIA ALGORITHMIC STABILITY

BY TIJANA ZRNIC[a] AND MICHAEL I. JORDAN[b]

Department of Electrical Engineering and Computer Sciences, University of California, Berkeley,
[a]*tijana.zrnic@berkeley.edu,* [b]*jordan@cs.berkeley.edu*

When the target of statistical inference is chosen in a data-driven manner, the guarantees provided by classical theories vanish. We propose a solution to the problem of inference after selection by building on the framework of *algorithmic stability*, in particular its branch with origins in the field of differential privacy. Stability is achieved via *randomization* of selection and it serves as a quantitative measure that is sufficient to obtain nontrivial post-selection corrections for classical confidence intervals. Importantly, the underpinnings of algorithmic stability translate directly into computational efficiency—our method computes simple corrections for selective inference without recourse to Markov chain Monte Carlo sampling.

1. Introduction. Classical statistical theory provides tools for valid inference under the assumption that the statistical question is determined before observing any data. In practice, however, the choice of question is typically guided by exploring the same data that is used for inference. This coupling between the statistical question and the data used for inference induces dependencies that invalidate guarantees derived from classical theories.

While traditional wisdom might deem this coupling unacceptable, recent literature embraces this coupled approach to statistical investigation and grants novel ways of thinking about validity. Indeed, data-driven model selection is widely taught and practiced, and even stands as a research area of its own. Sometimes model selection is even unavoidable; in the canonical setting of linear regression, the statistician often starts with a pool of candidate variables large enough that it makes the solution unidentifiable without additional constraints, and when those constraints are data-dependent the solution depends on the data in two ways.

This coupling of the problem formulation and inference stages of statistical analysis has been thoroughly studied in a line of work called *selective*, or *post-selection*, *inference* [7, 9, 47]. To this day, however, there are few general principles that enable both statistically powerful and computationally tractable inference after selection. Most existing solutions are either tailored to specific selection strategies (e.g., [30, 31, 53]), and as such do not generalize to all popular selection methods, or are valid for arbitrary selections at the cost of increased conservativeness (e.g., [2, 9]).

In the current paper, we build on concepts from the field of differential privacy [17, 18] to derive selective confidence intervals that are both tractable computationally and powerful statistically. Our theoretical framework delivers intervals of *tunable width*, a useful consequence of the fact that our confidence intervals derive from a quantitative measure of the *algorithmic stability* of the selection procedure. More precisely, we provide a valid correction to classical, nonselective confidence intervals simultaneously for all procedures that have the same level of algorithmic stability. Informally, a selection being stable means that it is not too sensitive to the particular realization of the data, and the more stable the selection is, the smaller the resulting intervals are. In particular, if the selection is "perfectly stable" in the sense that the

Received March 2022; revised May 2023.
MSC2020 subject classifications. Primary 62J15; secondary 62J05, 62F07.
Key words and phrases. Post-selection inference, selective inference, stability, differential privacy, model selection, linear regression.

inferential target is fixed up front and does not depend on the data at hand, the confidence intervals resulting from our approach smoothly recover classical confidence intervals.

We sketch our main result. Let $\hat{S}$ denote a data-dependent outcome of selection. For example, $\hat{S}$ could be subset of $\{1, \ldots, d\}$ corresponding to the variables selected for inclusion in a linear regression model. For every possible selection S, let β_S denote the resulting inferential target. In the linear regression context, β_S could be the population-level least-squares solution within the model determined by S.

Imagine that there is an oracle that guesses $\hat{S}$, only knowing the method used to arrive at the selection together with the distribution of the data, but not its realization. Denote by $\hat{S}_0$ the oracle's guess. We say that a selection procedure is η-stable for some $\eta > 0$ if there exists an oracle such that, with high probability over the distribution of the data, the likelihood of any selection under $\hat{S}$ and the likelihood of the same selection under $\hat{S}_0$ can differ by at most a multiplicative factor of e^{η}. Intuitively, η quantifies how much the selection can vary across different realizations of the data; $\eta = 0$ essentially means that the selection cannot depend on the data and hence $\hat{S}$ is fixed, while as η grows the selection is allowed to be increasingly data-adaptive. Note that the magnitude of stability depends not only on the selection method, but also on the distribution of the data.

Our main result provides a post-selection-valid correction to classical, nonselective confidence intervals for stable selection procedures. We state an informal version of our key inference tool.

THEOREM 1.1 (Informal). *For every* fixed *selection* S, *suppose that* $\mathrm{CI}_S^{(\alpha)}$ *are confidence intervals with valid coverage*,

$$\mathbb{P}\{\beta_S \notin \mathrm{CI}_S^{(\alpha)}\} \leq \alpha.$$

Let $\hat{S}$ *be an* η*-stable selection. Then*

$$\mathbb{P}\{\beta_{\hat{S}} \notin \mathrm{CI}_{\hat{S}}^{(\alpha e^{-\eta})}\} \leq \alpha.$$

Theorem 1.1 is valid simultaneously across *all* possible selection methods which are η-stable. In other words, under the computational notion of stability we consider, the stability parameter of a selection method alone is sufficient to correct for selective inferences.

Our stability designs are based on explicit randomization schemes which calibrate the level of randomization to a prespecified algorithmic stability requirement. Together with Theorem 1.1, this allows the statistician to *choose* the confidence interval width, obtaining a perturbation of a selection algorithm (e.g., the LASSO), to obtain a target interval width and a guarantee of valid coverage. Since the derived perturbation is an explicit function of the target interval width, this provides a way to understand the loss in utility due to randomization; for example, expressing how "far" the perturbed LASSO solution is from the standard, nonrandomized LASSO solution, in some appropriate sense. With this methodology in hand, one can explicitly analyze the inherent tradeoff between the post-selection correction and loss in utility due to randomization for any stable procedure.

We note that the use of randomization in selective inference is by no means a new idea (see, e.g., [10, 26, 36, 37, 49–51]). The main difference between our work and previous work is the use of stability as an analysis tool, which, on the one hand, leads to a computationally efficient, nonparametric, sampling-free approach to constructing selective confidence intervals with strict coverage, and on the other hand, explicitly connects the level of randomization to the resulting interval width. We elaborate on the comparisons to related work in Section 3.

1.1. *Organization.* In the following section, we present two motivating vignettes together with our solutions based on stability. In Section 3, we discuss related work. In Section 4, we introduce the notion of algorithmic stability at the focus of our study and in Section 5 we give theory for statistical inference under this definition. Then, in Section 6 we instantiate our theory in the context of model selection in linear regression. In Section 7, we draw connections to conditional post-selection inference. In Section 8, we discuss the design of stable algorithms and give stable versions of the LASSO and marginal screening. In Section 9, we study the performance of our procedures empirically. We end with a brief discussion in Section 10.

2. Motivating vignettes. To illustrate our framework, we present two motivating examples together with solutions implied by our theory, deferring the proofs of validity of the solutions to later sections. In addition, we compare our correction to some relevant baselines.

2.1. *Vignette* 1: *Winner's curse.* The first vignette considers the problem of selecting the largest observed effect. Suppose that we observe an n-dimensional vector $y \sim \mathcal{N}(\mu, \sigma^2 I)$; for example, each entry in this vector could be an observed treatment effect for a separate treatment. We are interested in doing inference on the most significant effect. More formally, denoting $i_* = \arg\max_i y_i$, we want to construct a confidence interval for μ_{i_*}. Note that this is a *random* inferential target because i_* is a function of the data.

One simple way of providing valid inference for μ_{i_*} is to apply the Bonferroni correction:

$$\mathbb{P}\{\mu_{i_*} \in (y_{i_*} \pm z_{1-\alpha/(2n)}\sigma)\} \geq 1-\alpha,$$

where z_q is the q quantile of the standard normal distribution.

Benjamini et al. [8] show that a tighter correction is valid, namely

$$\mathbb{P}\{\mu_{i_*} \in (y_{i_*} \pm z_{1-\alpha/(n+1)}\sigma)\} \geq 1-\alpha.$$

We show that, if we *randomize* the selection step, rather than select i_* exactly, the intervals can be made even tighter. Furthermore, the reduction in interval width is directly related to the amount of randomization.

CLAIM 1. *Suppose that we select* $\hat{i}_* = \arg\max_i (y_i + \xi_i)$, *where* $\xi_i \overset{\text{i.i.d.}}{\sim} \text{Lap}(\frac{2z_{1-\alpha\delta/(2n)}}{\eta})$, *for user-chosen parameters* $\eta > 0$, $\delta \in (0,1)$. *Then*

$$\mathbb{P}\{\mu_{\hat{i}_*} \in (y_{\hat{i}_*} \pm z_{1-\alpha(1-\delta)e^{-\eta/2}}\sigma)\} \geq 1-\alpha.$$

The proof of validity of this construction relies on our notion of stability, introduced in later sections; we defer the analysis of Claim 1 to Appendix A of the Supplementary Material [54]. Note that, as η and δ decrease toward zero, the noise level increases and the intervals approach classical, nonselective intervals. In general, η is the key parameter that trades off the information used for selection versus inference: small η corresponds to using more information for inference, while large η corresponds to prioritizing selection quality. Figure 1 illustrates how the interval width changes with η, in comparison to baselines, for $\sigma = 1$, $\delta = 0.5$, and varying n.

Note that if there is significant separation between μ_{i_*} and the other effects, $\hat{i}_*$ will likely be equal to i_* even for small η. If, on the other hand, there are multiple effects of similar magnitude, the randomization will smooth out the selection and place nonnegligible probability on all the competitive effects. In particular, to ensure $\hat{i}_* = i_*$ with high probability, it suffices to have η inversely proportional to the gap between the largest and second largest observed effect, $\Delta = y_{i_*} - \max_{j \neq i_*} y_j$.

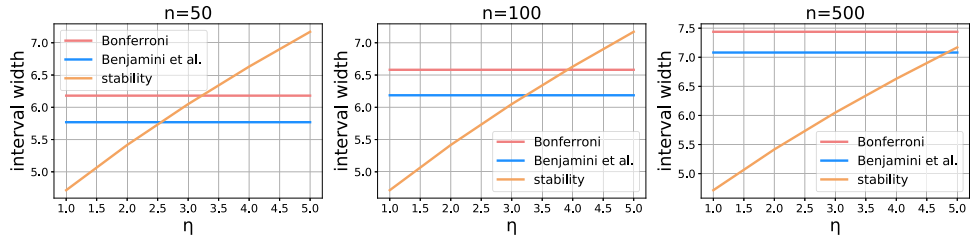

FIG. 1. *Confidence interval width around the "winning" effect, computed via the Bonferroni correction, Benjamini et al. [8] correction, and our stability-based approach. From left to right, we increase the value of* $n \in \{50, 100, 500\}$ *and keep* $\sigma = 1$ *fixed.*

CLAIM 2. *If* $\eta \geq \frac{4\log(n/(2\delta'))z_{1-\alpha\delta/(2n)}}{\Delta}$ *for some* $\delta' \in (0,1)$, *then* $\hat{i}_* = i_*$ *with probability at least* $1-\delta'$ *over the randomness in the selection.*

As a result, for large enough Δ, the approach of Claim 1 allows selecting i_* with high probability while providing a tighter correction than the baselines.

Finally, we note that, just like the Bonferroni correction, the stability-based solution can be applied nonparametrically. For example, the solution is applicable when the errors are only known to be sub-Gaussian, not necessarily Gaussian.

2.2. *Vignette* 2: *Feature selection.* In the second example, we look at inference after data-driven feature selection. Suppose we have a fixed design matrix, $X \in \mathbb{R}^{n\times d}$, with n observations and d features and a corresponding outcome vector $y \sim \mathcal{N}(\mu, \sigma^2 I) \in \mathbb{R}^n$. Denote by X_i the columns of X, for $i \in [d]$. We would like to select a *model* corresponding to a subset of the d features, and perform valid inference on the least-squares target after regressing y on the *selected* features only. This problem is discussed in depth by Berk et al. [9].

We set this problem up more generally in later sections; to keep this illustration light, assume that the features are normalized so that $\|X_i\|_2 = 1$ and we are selecting a single feature. Then, this problem amounts to doing inference on $X_{i_*}^\top \mu$, where i_* is the selected feature. Again, we note that this is a random inferential target since i_* is data-dependent.

Suppose that the goal of selection is to simply maximize the absolute correlation of the selected feature with y: $i_* = \arg\max_i |X_i^\top y|$. Then, our results imply the following.

CLAIM 3. *Suppose that we select* $\hat{i}_* = \arg\max_i |X_i^\top y + \xi_i|$, *where* $\xi_i \overset{\text{i.i.d.}}{\sim} \text{Lap}(\frac{2z_{1-\alpha\delta/(2d)}}{\eta})$, *for user-chosen parameters* $\eta > 0$, $\delta \in (0,1)$. *Then*

$$\mathbb{P}\{X_{\hat{i}_*}^\top \mu \in (X_{\hat{i}_*}^\top y \pm z_{1-\alpha(1-\delta)e^{-\eta}/2}\sigma)\} \geq 1-\alpha.$$

Again, as η and δ tend toward zero, the intervals approach nonselective intervals, and the relationship between η and the gap between the largest and second largest correlation, $\Delta = |X_{i_*}^\top y| - \max_{j\neq i_*} |X_j^\top y|$, drives the accuracy of selection.

CLAIM 4. *If* $\eta \geq \frac{4\log(d/(2\delta'))z_{1-\alpha\delta/(2d)}}{\Delta}$ *for some* $\delta' \in (0,1)$, *then* $\hat{i}_* = i_*$ *with probability at least* $1-\delta'$ *over the randomness in the selection.*

An alternative, equally simple solution to inference after feature selection is data splitting: we use a fraction $f \in (0,1)$ of the data for selection and the remaining $1-f$ fraction for inference. In Section 5.1, we present a more detailed comparison with data splitting. For now we remark that, given that both data splitting and stability lead to intervals that look like

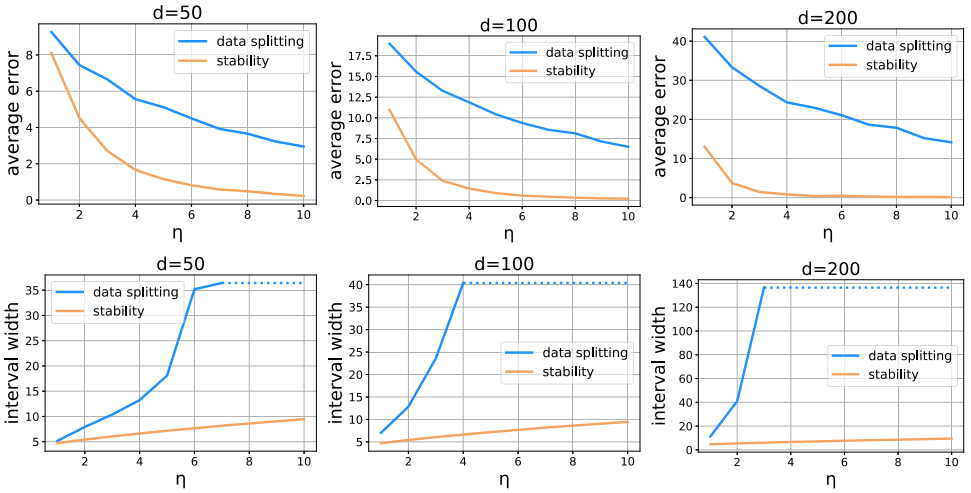

FIG. 2. *Average error, defined as* $\max_i |X_i^\top y| - |X_{\hat{i}_*}^\top y|$, *when interval widths of the two approaches are matched* (*top row*), *and average interval width when errors of the two approaches are matched* (*bottom row*). *We use one of the experimental setups in Section* 9. *The rows of* X *are drawn i.i.d. from an equicorrelated multivariate Gaussian distribution with pairwise correlation* 0.5. *The outcome is generated as* $y = X\beta + \epsilon$, *where* ϵ *is isotropic Gaussian noise and half of the entries in* β *are zero and half are sampled independently from* Exp(0.2). *From left to right, we increase the value of* $d \in \{50, 100, 200\}$ *and keep* $n = 50$ *fixed.*

classical intervals with an additional correction factor, every parameter η has a corresponding parameter $f \equiv f(\eta)$ that yields data-splitting intervals of roughly the same width as stability-based intervals (given a fixed δ such as 0.5). However, we observe that stability can be a significantly more powerful solution. The top row of Figure 2 compares the error of data splitting with the error of the stability solution described in Claim 3 in a simple simulation setting, when the interval widths implied by the two approaches are matched. In the bottom row of Figure 2 we provide a reverse comparison: we compare the average interval width of data splitting and the stability solution when the errors of the two approaches are matched. Unlike interval widths, we match the errors empirically: for each stability parameter η, we consider all splitting fractions f (equivalently, subsample sizes) and take the one that implies closest average error when averaged over 10,000 trials. Dotted lines indicate that the intervals are infinite (because the errors are matched when all data is used for selection, leaving no information for inference) or essentially infinite (larger than 200).

Finally, we emphasize that stability can be applied to the problem of feature selection even when data splitting is not an option, such as when there are spatial or temporal dependencies in the data.

3. Related work. In this section, we elaborate on the comparisons between our work and existing work in post-selection inference, and additionally discuss relevant work in the algorithmic stability literature.

3.1. *Simultaneous coverage.* In the formulation of post-selection inference by Berk et al. [9], the goal is to construct simultaneous confidence intervals (as per Eq. (2)) that are valid *for any model selection method* $\hat{M} : y \to \mathcal{M}$, for a prespecified model class $\mathcal{M}$. The framework of Berk et al. was subsequently generalized by Bachoc et al. [2] to handle distributions beyond the homoscedastic Gaussian, as initially assumed. These proposals are computationally infeasible in high dimensions as they essentially require looking for the "worst possible" model $M \in \mathcal{M}$, one that implies the largest so-called PoSI (Post-Selection Infer-

ence) constant, an analog of which we introduce and characterize in the context of stability. More recent work has proposed computationally efficient confidence regions via UPoSI [29].

Another approach to valid post-selection inferences that applies to general selection rules is *data splitting* [39]: split the data into two disjoint subsets, then use one subset to select the inferential target and the other subset to perform inference. Data splitting is appealing because, if the two subsets of the data are independent, classical inferences will be valid regardless of the selection procedure. However, data splitting is not universally applicable as one cannot always obtain two independent data sets, and even if applicable, it can suffer a significant loss in power, such as when only a few samples capture some relevant information. Our stability-based approach does not rely on any independence assumption between different observations, and, as illustrated in Figure 2, it can be a more powerful solution than data splitting when the latter is applicable. We give a further discussion of data splitting and its relationship to stability in Section 5.1.

All the aforementioned strategies strive for robustness: they protect against *all* selection procedures. For specific selection procedures, however, the intervals computed by simultaneous methods and related approaches are unnecessarily wide, as they do not exploit any knowledge of how the analyst arrives at the selected target. Recent work aims to address this issue in the framework of simultaneous coverage over the selected variables (SoS) by constructing SoS-controlling confidence intervals for k seemingly largest effects [8]. Our work likewise implies SoS intervals, by putting forward a general stability perspective and analyzing the relationship between stability and interval width for arbitrary stable procedures.

3.2. *Conditional coverage.* Conditional methods [21] exploit properties of the selection procedure. However, they control a different error criterion than simultaneous methods. In particular, the goal of conditional post-selection inference is to design CI_S such that, for all fixed selections S, $\mathbb{P}\{\beta_S \in \mathrm{CI}_S | \hat{S} = S\} \geq 1 - \alpha$. For a fixed selection procedure, conditional post-selection inference aims to characterize the distribution of the data given $\hat{S} = S$, and then using the knowledge of this conditional distribution it computes CI_S. This approach is tailored to the selection method at hand, and existing work has derived intervals for model selection via methods such as the LASSO [30], marginal screening, orthogonal matching pursuit [31], forward stepwise, and LARS [53].

It is often remarked that the conditional approach leads to *overconditioning*, thus leading to wide intervals [27]. Informally, overconditioning refers to the phenomenon of overstating the cost of selection, thus leaving little information for inference. Surprisingly, it has even been observed that simultaneous approaches can in some cases yield smaller intervals, due to the intervals being unconditional rather than conditional [2]. One attempt at narrowing down the intervals involves choosing a better event on which to condition [33]. Another solution to overconditioning which is relevant to the present context is the idea of randomizing the selection procedure [10, 26, 36, 37, 49–51]. Notably, the pioneering work in this direction due to Tian and Taylor [51] proves a central limit theorem that asymptotically relates the validity of statistical inferences without selection to their selective counterparts, a result similar in flavor to our Theorem 5.2. However, existing randomization proposals suffer several drawbacks. One is that they give little insight into the tradeoff between confidence interval width and the loss in utility from the additional noise. Another issue is that inference is based on a selective pivot which, unlike in exact conditional approaches, lacks closed-form expressions. As a result, to approximate the pivot, existing work resorts to computationally expensive sampling [49, 51], which is generally infeasible in high dimensions. There are other, computationally-efficient approaches which aim to approximate the pivot [36, 37], although these are only approximate and the general theory applies to restricted classes of selection problems.

Although our primary goal is to provide unconditional guarantees, in Section 7 we will also discuss implications of stability for conditional inference.

We also point out the work of Andrews et al. [1], who propose a hybrid approach that interpolates between unconditional and conditional post-selection inference to obtain smaller confidence intervals relative to a purely conditional approach.

3.3. *Algorithmic stability.* The technical tools of this paper are rooted in the theory of differential privacy [17, 18] and its extensions [4, 15]. Initially, differential privacy was developed as a standard for private data analysis. A more recent line of work, typically referred to as *adaptive data analysis* (see, e.g., [5, 15, 16]), has recognized that a stability concept can be extracted from differential privacy and exploited to obtain perturbation-based generalization guarantees in learning theory. Superficially, adaptive data analysis has the same goal as post-selection inference—developing statistical tools for valid inference when hypotheses about the data are also data-driven—but the typical formalization of this problem is not directly comparable to that of the canonical post-selection inference setup in regression. The conceptual connection between the two areas has, however, been recognized (dating back to at least the seminal work of Tian and Taylor [51]), and several existing works discuss selective hypothesis tests and stability-based corrections to arbitrary selective p-values [40, 42]. Our work does not aim to contribute to adaptive data analysis per se; rather, we build on and adapt existing tools in this literature for the purpose of providing post-selection corrections for common selection problems, such as within the framework put forward by Berk et al. [9]. Finally, we note that connections between stability and generalization are not new [11], and stability ideas have been utilized to construct predictive confidence intervals [3, 44].

4. Algorithmic stability and selection. The formal theory of algorithmic stability characterizes how the output of an algorithm changes when the input is perturbed. Randomized algorithms have as output a *random variable*; therefore, to study the stability of a randomized algorithm, an appropriate notion of closeness of two random variables is required. The particular notion of closeness considered in differential privacy and related work is known as *indistinguishability*, or *max-divergence*.

DEFINITION 4.1 (Indistinguishability). A random variable Q is (η, τ)-indistinguishable from W, denoted $Q \approx_{\eta,\tau} W$, if for all measurable sets $\mathcal{O}$, $\mathbb{P}\{Q \in \mathcal{O}\} \leq e^{\eta}\mathbb{P}\{W \in \mathcal{O}\} + \tau$.

Note that indistinguishability is essentially a property of two distributions; for this reason, we will sometimes say that a distribution $\mathcal{P}_Q$ is (η, τ)-indistinguishable from a distribution $\mathcal{P}_W$, meaning that $Q \approx_{\eta,\tau} W$ holds for any $Q \sim \mathcal{P}_Q$ and $W \sim \mathcal{P}_W$.

Roughly speaking, τ bounds the probability of the event where Q and W are "very different." For fixed $\tau \in [0, 1]$, the parameter η is meant to capture how similar the distributions of Q and W are—the larger η is the larger the divergence between Q and W can be. One should think of τ as being at most a small factor proportional to the miscoverage probability α.

We now formally introduce the main notion of algorithmic stability considered in this paper. The algorithm whose stability we analyze will usually be a selection algorithm. Intuitively, a randomized algorithm $\mathcal{A}$ is stable if there exists an "oracle" random variable A_0 such that, for all "typical" inputs ω, $\mathcal{A}(\omega)$ is distributionally indistinguishable from A_0. In other words, as long as the input is typical, we can approximate the distribution of the randomized algorithm's output with a *fixed* law, without having to see the input in the first place.

DEFINITION 4.2 (Stability). Let $\mathcal{A} : \mathbb{R}^n \to \mathcal{S}$ be a randomized algorithm. We say that $\mathcal{A}$ is (η, τ, ν)-stable with respect to a distribution $\mathcal{P}$ supported on $\mathbb{R}^n$ if there exists a random variable A_0, possibly dependent on $\mathcal{P}$, such that $\mathcal{P}\{\omega \in \mathbb{R}^n : \mathcal{A}(\omega) \approx_{\eta,\tau} A_0\} \geq 1 - \nu$.

This notion is a special case of *typical stability* introduced by Bassily and Freund [4]. It is closely related to the notions of perfect generalization [13] and max-information [15]. Unless stated otherwise, whenever we use the term stability we will assume stability in the sense of Definition 4.2. The parameter ν can in principle take on any value in $[0, 1]$ but in practice we will set it to be proportional to α.

We will only invoke stability with respect to the data distribution, which we will denote by $\mathcal{P}_y$. Thus, for simplicity, when we say that $\mathcal{A}$ is (η, τ, ν)-stable we are implicitly assuming that it is stable with respect to $\mathcal{P}_y$.

Definition 4.2 requires that, as the input data ω varies, the distribution of $\mathcal{A}(\omega)$ remains indistinguishable from a *fixed* distribution that does not depend on ω, namely the distribution of A_0. The parameter ν allows the laws of $\mathcal{A}(\omega)$ and A_0 to deviate for a small set of atypical data vectors ω. The parameters η and τ bound the maximum deviation of $\mathcal{A}(\omega)$ from A_0 over the typical set of vectors ω.

Given a stable algorithm, we will refer to A_0 (which must exist by definition) as its corresponding oracle. The term "oracle" is motivated by the fact that A_0 will typically depend on $\mathcal{P}_y$, which is unknown. To build further intuition, suppose that we observe data $y \sim \mathcal{P}_y$ and let $\mu = \mathbb{E}y$. Most of our stability constructions will rely on arguing that Definition 4.2 holds if we take $A_0 = \mathcal{A}(\mu)$; the reader should think of this as the most prototypical oracle construction. In other words, $\mathcal{A}(y)$ *conditional* on y is indistinguishable from $A(\mu)$ in the sense of Definition 4.1 (as long as y is not an atypical data set). At a high level, this happens because y concentrates around μ; we work out a concrete example building on this idea below.

EXAMPLE. To provide intuition for Definition 4.2, we present one simple mechanism for achieving stability. Although basic, this mechanism will be a fundamental building block in our stability proofs. Suppose that we wish to compute $w^\top y$, for some fixed vector w, and suppose that we take $\mathcal{P}_y$ to be $\mathcal{N}(\mu, \sigma^2 I)$ with known $\sigma > 0$. Let $\mathcal{A}(y) = w^\top y + \xi$, where $\xi \sim \mathrm{Lap}(\frac{z_{1-\nu/2}\sigma\|w\|_2}{\eta})$, for user-specified parameters $\eta > 0$, $\nu \in (0, 1)$. Here, $\mathrm{Lap}(b)$ denotes a draw from the zero-mean Laplace distribution with parameter b, independent of y. We argue that this mechanism is $(\eta, 0, \nu)$-stable. First, we know

$$\mathbb{P}\{|w^\top y - w^\top \mu| \geq z_{1-\nu/2}\sigma\|w\|_2\} = \mathbb{P}\{|\mathcal{N}(0, \sigma^2\|w\|_2^2)| \geq z_{1-\nu/2}\sigma\|w\|_2\} = \nu.$$

Denote $E = \{\omega \in \mathbb{R}^n : |w^\top \omega - w^\top \mu| \leq z_{1-\nu/2}\sigma\|w\|_2\}$, and notice that we have shown that $\mathbb{P}\{y \in E\} = 1 - \nu$.

Now let $A_0 = \mathcal{A}(\mu)$. Since the ratio of densities of $\xi \sim \mathrm{Lap}(b)$ and its shifted counterpart $x + \xi$ is upper bounded by $e^{|x|/b}$, we can conclude that for all $\omega \in E$ and measurable sets $\mathcal{O}$,

$$\frac{\mathbb{P}\{\mathcal{A}(\omega) \in \mathcal{O}\}}{\mathbb{P}\{\mathcal{A}(\mu) \in \mathcal{O}\}} \leq e^{\eta};$$

that is, we have $\mathcal{A}(\omega) \approx_{\eta,0} A_0$ for all $\omega \in E$. Putting everything together, we see that $\mathcal{A}(\cdot)$ is $(\eta, 0, \nu)$-stable with respect to $\mathcal{P}_y$.

Throughout we will use $\hat{S}(\cdot)$ to denote a possibly randomized *selection* algorithm, which takes as input the data y and outputs a selection that determines the inferential target. For example, $\hat{S}$ could be a model selection algorithm such as in the second vignette, or it could be an algorithm that selects an effect that is the focus of subsequent inference, such as in the first vignette. With a slight abuse of notation, we will use $\hat{S}$ to denote both the mapping from the data to the selection as well as the selection itself, $\hat{S}(y) \equiv \hat{S}$.

5. Confidence intervals after stable selection. Given the assumption of (η, τ, ν)-stability, we now show how a simple modification to classical confidence intervals suffices to correct for selective inferences. This correction is valid *regardless* of any additional property of the selection criterion.

The main intuition behind this assertion is the following. If the selection algorithm is stable, then by Definition 4.2 one can construct an oracle selection $\hat{S}_0$ *without looking at* y, such that $\hat{S}(y)$ and $\hat{S}_0$ are distributionally indistinguishable. Since $\hat{S}(y)$ is indistinguishable from $\hat{S}_0$, we can pretend that $\hat{S}_0$ is the selection of interest. Furthermore, since $\hat{S}_0$ was constructed independently of y, we are *free to use* y *for inference*. Stability ensures that, despite data reuse, inference behaves almost like with data splitting, in which we perform selection on one batch of data and then use independent data for constructing intervals.

We state a technical lemma, related to Lemma 3.3 by Bassily and Freund [4], that we use to prove our main theorem. We include a proof of Lemma 5.1 in Appendix A of the Supplementary Material [54].

LEMMA 5.1. *Let* $\hat{S} : \mathbb{R}^n \to \mathcal{S}$ *be an* (η, τ, ν)*-stable selection algorithm and let* $\hat{S}_0$ *be the corresponding oracle selection. Then, it holds that* $(y, \hat{S}(y)) \approx_{\eta, \tau+\nu} (y, \hat{S}_0)$.

Equipped with Lemma 5.1, we can now describe how to construct post-selection-valid confidence intervals after stable selection.

Suppose that, under selection S, our target of inference is β_S. Moreover, suppose that $\mathrm{CI}_S^{(\alpha)}$ are valid intervals at level $1 - \alpha$ for any *fixed* S, meaning $\mathbb{P}\{\beta_S \notin \mathrm{CI}_S^{(\alpha)}\} \leq \alpha$. Such intervals are provided by classical theory.

Theorem 5.2 formally states how to construct confidence intervals for an *adaptive* target $\beta_{\hat{S}}$, when $\hat{S}$ is selected in a stable way. This is the key result of our paper.

THEOREM 5.2. *Fix* $\delta \in (0, 1)$, *and let* $\hat{S}$ *be an* (η, τ, ν)*-stable selection algorithm. Then,*

$$\mathbb{P}\{\beta_{\hat{S}} \notin \mathrm{CI}_{\hat{S}}^{(\delta e^{-\eta})}\} \leq \delta + \tau + \nu.$$

In words, if $\hat{S}$ is (η, τ, ν)-stable, we can pretend that there is no selection bias and simply construct classical intervals, albeit at a more conservative level, to achieve validity. If we set the target error level to be $\delta e^{-\eta}$, then the realized error level will be at most $\delta + \tau + \nu$.

5.1. *Comparison with data splitting.* In many scenarios it is possible to split the data into two independent chunks, one to be used for selection and the other to be reserved for inference. Classical inferences are then valid because the inferential target is determined before seeing any of the data used in the inference step. This simple baseline for valid inference after selection is called *data splitting*. In this section, we illuminate the relationship between our approach via stability and data splitting.

First we want to emphasize that the stability principle is applicable even with dependent samples: Theorem 5.2 can be applied even when it is not clear how to create two independent subsets of the data. Moreover, in some selection problems data splitting makes little conceptual sense, such as in our first motivating vignette about inference on the winning effect.

The appeal of data splitting lies in its broad applicability. As long as the data can be split into two independent components, the criteria for choosing the inferential target can be arbitrary. Therefore, data splitting provides a *selection-agnostic* correction, universally valid across all possible selection strategies.

Conceptually, stability lies somewhere between data splitting and conditional post-selection inference. It computes a correction level as a function of how adaptive the selection

is to the data, thereby adapting to some properties of the selection rule like conditional inference methods. However, at the same time it provides a correction that is universally valid across all possible selection strategies *with the same level of stability*, which can be seen as a refinement of the principle of data splitting.

To illustrate the conceptual difference between the stability principle and the data splitting principle, suppose that in the latter case we allocate f-fraction of the data to selection, and $(1-f)$-fraction to inference. Then, the resulting intervals will roughly look like classical intervals augmented by a factor of $\sqrt{\frac{1}{1-f}}$ *regardless* of how the selection is performed.

In contrast, the stability approach augments classical intervals as a function of the adaptivity of the selection algorithm. Suppose for concreteness that $y \sim \mathcal{N}(\mu, I)$ and we are considering doing inference on one of two targets, $v_0^\top \mu$ or $v_1^\top \mu$, where the selection $\hat{S} \in \{0,1\}$ depends on the data y. Consider three different selection methods:

- $\hat{S} = 1$ no matter what the data vector is.
- $\hat{S} = 1$ if $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i \geq 0$, and $\hat{S} = 0$ otherwise.
- $\hat{S} = 1$ if $X_1^\top y \geq 0$ for some unit vector X_1, and $\hat{S} = 0$ otherwise.

We can write all three procedures as $\hat{S} = \mathbf{1}\{w^\top y \geq 0\}$; in the first case $w = 0$, in the second case $w = \frac{1}{n}\mathbf{1}$, and in the third case $w = X_1$.

Let us fix the noise level $\gamma > 0$ and select $\hat{S} = \mathbf{1}\{w^\top y + \xi \geq 0\}$, where $\xi \sim \mathrm{Lap}(\gamma)$. The first method is trivially $(0,0,0)$-stable for any level γ, hence we can simply use y for inference without any correction. Based on the same analysis as in the example in Section 4, the second selection method is $(\sqrt{2\log(2/\nu)}/(\gamma\sqrt{n}), 0, \nu)$-stable for all $\nu > 0$; that is, $(\sqrt{2\log(4/\alpha)}/(\gamma\sqrt{n}), 0, \alpha/2)$-stable. Similarly, the third selection method is $(\sqrt{2\log(4/\alpha)}/\gamma, 0, \alpha/2)$-stable.

We can thus observe that, even though in all three examples we perturb the selection by the same constant level of noise, the stability approach exploits the fact that some selection criteria are more stable than others and this is reflected in the resulting stability parameter. By Theorem 5.2, this stability parameter, in turn, directly determines the correction factor, that is, how conservative we need to make classical inferences for them to be valid post selection.

While data splitting and stability come with conceptual differences, they also have technical similarities. In particular, each one has a leading parameter—$f \in (0,1)$ in the case of data splitting and $\eta > 0$ in the case of stability—and this parameter interpolates between two extremes. One extreme is when all information is reserved for inference (attained when $f = 0$ and $\eta = 0$, respectively) and the other is when all information is used for selection (attained when $f = 1$ and $\eta \to \infty$, respectively). Therefore, it might make sense to ask how the two interpolations relate.

For every η, there is an $f(\eta)$ such that, if we used $f(\eta)$-fraction of the data for selection and $1 - f(\eta)$ for inference, we would approximately get the same interval correction. We sketch the derivation of $f(\eta)$ in the case of normal intervals for simplicity, however this calculation can be generalized to other distributions. We will assume that $\nu + \tau \leq \delta\alpha$ for some $\delta \in (0,1)$; then, the intervals resulting from (η, τ, ν)-stability are of width proportional to $z_{1-(1-\delta)\frac{\alpha}{2}e^{-\eta}}$. The intervals resulting from data splitting are of width proportional to $z_{1-\frac{\alpha}{2}}(1-f(\eta))^{-1/2}$. By equating the two expressions to achieve the same width and simplifying, we obtain

$$f(\eta) = 1 - \left(\frac{z_{1-\frac{\alpha}{2}}}{z_{1-(1-\delta)\frac{\alpha}{2}e^{-\eta}}}\right)^2 \approx \frac{\log\frac{1}{1-\delta} + \eta}{\log\frac{2}{(1-\delta)\alpha} + \eta}, \tag{1}$$

where the approximation on the right-hand side follows by a sub-Gaussian approximation.

Of course, this sketch only gives intuition for when data splitting and stability imply equally powerful inference; it does not say anything about which selection is more accurate—one where we select on $f(\eta)$-fraction of the data, or one where we select on the whole data set in an η-stable way. We will tackle this question empirically, as the notion of "more accurate" varies greatly depending on the context. In Figure 2, we used the splitting fraction in Eq. (1) and observed that stability outperforms data splitting. We provide further empirical comparisons in Section 9.

Finally, we mention another proposal that is conceptually closely related to data splitting, namely the (U, V) decomposition of Rasines and Young [38]. Like stability, the (U, V) decomposition allows the statistician to see all data points—more precisely, noisy versions thereof—both in the selection step and in the inference step. This is an important advantage over data splitting when there are only a few samples that capture information about certain directions. In contrast with stability, performing the (U, V) decomposition does not rely on any properties of the selection method. However, finite-sample guarantees of this approach crucially rely on the data being Gaussian with known covariance, while the stability principle is applicable beyond Gaussianity and is robust to only having an estimate of the covariance.

6. Model selection in linear regression. In this section, we discuss an application of our stability tools to the problem of model selection in linear regression. We focus on the framework presented in the seminal work of Berk et al. [9]. We begin by reviewing the model and introduce the necessary notation.

Let $X \in \mathbb{R}^{n\times d}$ denote a fixed design matrix, and let $X_i \in \mathbb{R}^n$ denote the ith column of X, for $i \in [d]$. We refer to vectors X_i as variables or features. For a subset $M \subseteq [d]$, we denote by $X_M \in \mathbb{R}^{n\times|M|}$ the submatrix of X given by selecting the columns indexed by M. We make no assumptions about how n and d relate; in particular, we could have $d \gg n$.

By $y \in \mathbb{R}^n$ we denote the random vector of outcomes corresponding to X. Importantly, we do not assume knowledge of a true data-generating process; for example, we do not assume that $\mu := \mathbb{E}[y]$ can be expressed as a linear combination of $\{X_i\}_{i=1}^d$. The vector $\mu \in \mathbb{R}^n$ is unconstrained and need not reside in the column space of X. Rather, different subsets of $\{X_i\}_{i=1}^d$ provide different approximations to μ, some better than others.

The statistician wishes to let the data decide how the initial pool of features should be reduced to a smaller set of seemingly relevant features, and then run linear regression on this smaller set. That is, the statistician chooses a set $\hat{M} \subseteq [d]$ by running a model selection method on X, y, and then aims to approximate $y \approx X_{\hat{M}}\hat{\beta}_{\hat{M}}$, for some $\hat{\beta}_{\hat{M}}$. As before, we will employ a conventional abuse of notation by letting $\hat{M} \equiv \hat{M}(y)$.

Assuming $X_{\hat{M}}$ has full column rank almost surely, the unique least-squares estimate in model $\hat{M}$ is given by

$$\hat{\beta}_{\hat{M}} := \underset{\beta\in\mathbb{R}^{|\hat{M}|}}{\arg\min} \|y - X_{\hat{M}}\beta\|_2^2 = \big(X_{\hat{M}}^\top X_{\hat{M}}\big)^{-1} X_{\hat{M}}^\top y := X_{\hat{M}}^+ y,$$

where we define $X_{\hat{M}}^+ := (X_{\hat{M}}^\top X_{\hat{M}})^{-1} X_{\hat{M}}^\top$ to be the pseudoinverse of $X_{\hat{M}}$. For a *fixed* model M, the target estimand of $\hat{\beta}_M$ is

$$\beta_M := \underset{\beta\in\mathbb{R}^{|M|}}{\arg\min}\, \mathbb{E}\big[\|y - X_M\beta\|_2^2\big] = X_M^+\mu,$$

and hence for a random model $\hat{M}$, this implies a *random* target $\beta_{\hat{M}} = X_{\hat{M}}^+\mu$.

We denote by $\beta_{j\cdot M}$ the entry of β_M corresponding to feature X_j, for all $j \in M$. Note that $\beta_{j\cdot M}$ is not defined for $j \notin M$. We adopt similar notation for the entries of $\hat{\beta}_M$.

Our goal is to construct simultaneous confidence intervals for the target of inference $\beta_{\hat{M}}$. More precisely, we wish to design $\mathrm{CI}^{(\alpha)}_{j\cdot\hat{M}}$ such that

$$\mathbb{P}\{\beta_{j\cdot\hat{M}} \in \mathrm{CI}^{(\alpha)}_{j\cdot\hat{M}},\ \forall j \in \hat{M}\} \geq 1-\alpha, \tag{2}$$

for a fixed $\alpha \in (0,1)$ and a *fixed* selection procedure $\hat{M}$. Note that the work of Berk et al. and various extensions [2, 9, 29] provide simultaneity *both* over the selected variables *and* over all selection methods, while we keep the selection method fixed. Our guarantees are *simultaneous over the selected* (cf. [8]).

The intervals resulting from our approach are of the form $\mathrm{CI}_{j\cdot\hat{M}}(K) := (\hat{\beta}_{j\cdot\hat{M}} \pm K\hat{\sigma}_{j\cdot\hat{M}})$, where $\hat{\sigma}^2_{j\cdot\hat{M}}$ is an estimator of variance for the OLS estimate $\hat{\beta}_{j\cdot\hat{M}}$; for example, the "sandwich" variance estimator [12]. Our goal is to find a suitable value of K such that $\mathrm{CI}_{j\cdot\hat{M}}(K)$ are valid $(1-\alpha)$-confidence intervals, as per Eq. (2). By analogy with Berk et al. [9], we refer to the minimal such valid K as the *PoSI constant*. It is important to remember that, unlike in Berk et al., our PoSI constant depends on the selection procedure, rather than a family of all possible models.

The PoSI constant is well characterized when the model is fixed rather than determined in a data-driven fashion. For a fixed model M and given $\alpha \in (0,1)$, we define $K_{M,\alpha}$ to be the minimum value of K such that

$$\mathbb{P}\left\{\max_{j\in M}\left|\frac{\hat{\beta}_{j\cdot M}-\beta_{j\cdot M}}{\hat{\sigma}_{j\cdot M}}\right| \geq K\right\} \leq \alpha.$$

In other words, $K_{M,\alpha}$ defines the PoSI constant when the model M is specified up front and does not depend on the data; in this case, $\mathrm{CI}_{j\cdot M}(K_{M,\alpha})$ are valid simultaneous intervals at level $1-\alpha$. For example, when $y \sim \mathcal{N}(\mu, \sigma^2 I)$, one simple way of providing a valid upper bound on $K_{M,\alpha}$ is via standard z-scores or t-scores, after doing a Bonferroni correction over $j \in M$. Sharper estimates of $K_{M,\alpha}$ can be obtained by exploiting the correlations between the regression coefficients to estimate the maximum z-score or t-score. Even in a distribution-free setting, it is common to determine $K_{M,\alpha}$ via normal approximation [35, 39].

We are now ready to state a corollary of Theorem 5.2 that focuses on the problem of model selection in linear regression.

COROLLARY 6.1. *Fix $\delta \in (0,1)$. Let $\hat{M}$ be an (η, τ, ν)-stable model selection algorithm. For all $j \in \hat{M}$, let $\mathrm{CI}_{j\cdot\hat{M}}(K_{\hat{M},\delta e^{-\eta}}) = (\hat{\beta}_{j\cdot\hat{M}} \pm K_{\hat{M},\delta e^{-\eta}}\hat{\sigma}_{j\cdot\hat{M}})$. Then*

$$\mathbb{P}\{\exists j \in \hat{M} : \beta_{j\cdot\hat{M}} \notin \mathrm{CI}_{j\cdot\hat{M}}(K_{\hat{M},\delta e^{-\eta}})\} \leq \delta + \tau + \nu.$$

To provide further intuition, we instantiate Corollary 6.1 in the canonical setting of Gaussian observations. Let $y \sim \mathcal{N}(\mu, \sigma^2 I)$. If $\sigma > 0$ is known, we let $\hat{\sigma}_{j\cdot M} = \sigma\sqrt{((X_M^\top X_M)^{-1})_{jj}}$; otherwise, we assume we have access to an estimate of σ, denoted $\hat{\sigma}$, and let $\hat{\sigma}_{j\cdot M} = \hat{\sigma}\sqrt{((X_M^\top X_M)^{-1})_{jj}}$. Following the treatment of Berk et al. [9], we assume that $\hat{\sigma}^2 \sim \sigma^2\frac{\chi^2_r}{r}$ for r degrees of freedom and assume that $\hat{\sigma}^2 \perp \hat{\beta}_{j\cdot M}$ for all possible OLS estimates $\hat{\beta}_{j\cdot M}$. If the full model is assumed to be correct, that is $y \sim \mathcal{N}(X\beta, \sigma^2 I)$, and $n > d$, then this assumption is satisfied for $r = n-d$ by setting $\hat{\sigma}^2 = \|y - X\hat{\beta}\|_2^2/(n-d)$, where $\hat{\beta}$ is the OLS estimate in the full model. Even if the full model is not correct, there exist other ways of producing a valid estimate of σ; we refer the reader to Berk et al. [9] for further discussion.

We denote by $z_{1-\alpha}$ the $1-\alpha$ quantile of the standard normal distribution, and by $t_{r,1-\alpha}$ the $1-\alpha$ quantile of the t-distribution with r degrees of freedom.

COROLLARY 6.2. *Fix $\delta \in (0,1)$, and suppose $y \sim \mathcal{N}(\mu, \sigma^2 I)$. Further, let $\hat{M}$ be an (η, τ, ν)-stable model selection algorithm. If σ is known, let*

$$\mathrm{CI}_{j\cdot\hat{M}} = \big(\hat{\beta}_{j\cdot\hat{M}} \pm z_{1-\delta/(2|\hat{M}|e^{\eta})}\sigma\sqrt{\big((X_{\hat{M}}^{\top}X_{\hat{M}})^{-1}\big)_{jj}}\big).$$

If, on the other hand, σ is not known but there exists an estimate, $\hat{\sigma}^2 \sim \sigma^2\frac{\chi_r^2}{r}$, independent of the OLS estimates, let

$$\mathrm{CI}_{j\cdot\hat{M}} = \big(\hat{\beta}_{j\cdot\hat{M}} \pm t_{r,1-\delta/(2|\hat{M}|e^{\eta})}\hat{\sigma}\sqrt{\big((X_{\hat{M}}^{\top}X_{\hat{M}})^{-1}\big)_{jj}}\big).$$

In either case, we have $\mathbb{P}\{\exists j \in \hat{M} : \beta_{j\cdot\hat{M}} \notin \mathrm{CI}_{j\cdot\hat{M}}\} \le \delta + \tau + \nu$.

The proof follows by a direct application of Corollary 6.1, together with a Bonferroni correction over $j \in \hat{M}$ when computing $K_{\hat{M},\delta e^{-\eta}}$. Approximating Gaussian quantiles by sub-Gaussian concentration, we observe that the PoSI constant in Corollary 6.2 scales roughly as $\sqrt{2(\log(2|\hat{M}|/\delta) + \eta)}$ (when σ is known, or as $r \to \infty$ when σ is estimated from data).

6.1. *Recovering the Scheffé rate.* Our main technical step in deriving selective confidence intervals is Lemma 5.1, which argues that the joint distribution of $(y, \hat{S})$ cannot be too different from the joint distribution of $(y, \hat{S}_0)$, where $\hat{S}_0$ is the oracle from the definition of stability. In the context of model selection in linear regression, we verify that the confidence intervals resulting from this approach are not vacuously wide in the two most extreme settings: the first, in which the model selection is independent of the data, and the second, in which the model selection is arbitrarily complex and dependent on the data.

Suppose that $\hat{M}$ is independent of y. Then, the distribution of $\hat{M}(y)$, conditional on y, is equal to the distribution of $\hat{M}(\omega)$ for any point ω, hence $\hat{M}(\omega)$ is an oracle which trivially implies $(0,0,0)$-stability. In this case, the intervals in Corollary 6.1 reduce to $\mathrm{CI}_{j\cdot\hat{M}}(K_{\hat{M},\delta})$ and are valid at level $1-\delta$, as expected.

Now suppose that $\hat{M}$ is allowed to have arbitrary dependence on y; in particular, it can attain the "significant triviality bound" of Berk et al. [9]. While arguing stability in the sense of Definition 4.2 would require additional assumptions, the only property of stability used to prove Theorem 5.2—the indistinguishability in Lemma 5.1—can be obtained. This allows for the proof of Theorem 5.2 to go through, thus recovering the tight rate of existing analyses.

PROPOSITION 6.3. *Let $\hat{M}$ be an arbitrary, possibly randomized model selection procedure, such that $|\hat{M}| \le s$ almost surely. Then, for any $\mathcal{P}_y$, there exists an oracle selection $\hat{M}_0$ such that for any $\tau \in (0,1)$,*

$$\big(y, \hat{M}(y)\big) \approx_{\eta,\tau} (y, \hat{M}_0), \quad \text{for some } \eta = O\big(s\log(d/s)\big) + \log(1/\tau).$$

Consequently, there exists a value $\eta = O(s\log(d/s)) + \log(1/\tau)$ such that the confidence intervals $\mathrm{CI}_{j\cdot\hat{M}}(K_{\hat{M},\delta e^{-\eta}}) = (\hat{\beta}_{j\cdot\hat{M}} \pm K_{\hat{M},\delta e^{-\eta}}\hat{\sigma}_{j\cdot\hat{M}})$ satisfy

$$\mathbb{P}\big\{\exists j \in \hat{M} : \beta_{j\cdot\hat{M}} \notin \mathrm{CI}_{j\cdot\hat{M}}(K_{\hat{M},\delta e^{-\eta}})\big\} \le \delta + \tau.$$

By approximating Gaussian quantiles via sub-Gaussian concentration, we obtain confidence intervals which are universally valid for *all* s-sparse selections under Gaussian outcomes and scale as $O(\sqrt{\eta}) = O(\sqrt{s\log(d/s)})$. This rate is in general tight [28], and as s approaches d, it matches the rate given by the Scheffé protection [9, 43].

7. Conditional coverage. So far all results we have presented have been about marginal coverage. Sometimes it is desirable to provide *conditional* coverage, whereby we condition on the event that a given selection was made. We discuss how stability can provide guarantees that closely resemble those of conditional post-selection inference.

If a selection $\hat{S}$ is (η, τ, ν)-stable, we know that there must exist an oracle selection $\hat{S}_0$ such that $\mathcal{P}_y\{\omega : \hat{S}(\omega) \approx_{\eta,\tau} \hat{S}_0\} \geq 1 - \nu$. In what follows, let E denote the set over which $\hat{S}$ is indistinguishable from the corresponding oracle $\hat{S}_0$: $E = \{\omega : \hat{S}(\omega) \approx_{\eta,\tau} \hat{S}_0\}$. Note that we know $\mathbb{P}\{y \in E\} \geq 1 - \nu$ by definition.

Importantly, the set E is known to the analyst. The reason is that E is specified as part of a stable algorithm design: to be able to claim that an algorithm is stable—meaning, indistinguishable from an oracle $\hat{S}_0$—one must provide a set E where $\hat{S}(\omega)$ is indistinguishable from $\hat{S}_0$ for all $\omega \in E$. To give an example, in the first motivating vignette in Section 2, E is taken to be all vectors ω such that $\|\omega - \mu\|_\infty$ is small (for an appropriately chosen radius depending on ν). For a desired stability level η, the magnitude of randomization is then calibrated to the size of this radius; that is, the size of E.

We state an implication of Lemma 5.1, the key step toward a conditional guarantee.

LEMMA 7.1. *Suppose that $\hat{S}$ is $(\eta, 0, \nu)$-stable with respect to oracle $\hat{S}_0 = \hat{S}(y'_E)$, where y'_E is a sample from $\mathcal{P}_y$ truncated to E. Then, it holds that*

$$\mathbb{P}\big\{y \in \mathcal{O}_S | \hat{S}(y) = S, y \in E\big\} \leq e^{\eta}\mathbb{P}\{y \in \mathcal{O}_S | y \in E\}, \tag{3}$$

for all selections S and measurable sets $\mathcal{O}_S$.

As suggested by Lemma 7.1, the main difference between conditional post-selection inference and the conditional guarantees implied by stability is that in the latter case we additionally truncate the distribution of y to a high-probability set E. Note that on the right-hand side of Eq. (3) there is no dependence on the selection event, which makes inference, despite selection, essentially as easy as classical inference.

We illustrate the conditional properties of stability with an example.

7.1. *Example*: *Publication bias*. We consider an illustration of the publication bias problem, also known as the file-drawer problem [21, 41, 51]. Suppose we observe an effect $y \sim \mathcal{P}_y$ with $\mathbb{E}[y] = \mu$, $\operatorname{supp}(\mathcal{P}_y) \subseteq \mathbb{R}$. We are interested in constructing an interval for μ only if the observed effect is deemed "interesting" enough, for example, if $y > T$ for some threshold T. Denote by $\text{report}(y)$ the event that we decide to report the confidence interval.

One approach to this problem is to evaluate the distribution of the data *conditional* on the selection event. For example, we could find K such that $\mathbb{P}\{|y - \mu| > K | \text{report}(y)\} \leq \alpha$, and report $\text{CI}(K) = (y \pm K)$ on the event $\text{report}(y)$. Importantly, this approach generally requires an explicit characterization of the event $\text{report}(y)$. Our theory suggests a *criterion-agnostic* solution based on randomizing the selection.

CLAIM 5. *Let $y \sim \mathcal{N}(\mu, \sigma^2)$. Suppose that we apply the selection criterion to $y + \xi$, where $\xi \sim \text{Lap}(b)$ for some user-chosen parameter $b > 0$; that is, we report the confidence interval on the event* $\text{report}(y + \xi)$. *Then, for any user-chosen parameter $\nu \in (0, 1)$, we have*

$$\mathbb{P}\big\{\mu \notin (y \pm z_{1-\frac{\alpha}{2}(1-\nu)e^{-\eta}}\sigma) | \text{report}(y + \xi), y \in E\big\} \leq \alpha,$$

where

$$\eta = \frac{z_{1-\nu/2}\sigma}{b} - \frac{\sigma^2}{2b^2} + \log\Big(\frac{1-\nu}{2(\Phi(z_{1-\nu/2} + \frac{\sigma}{b}) - \Phi(\frac{\sigma}{b}))}\Big)$$

and E is an event such that $\mathbb{P}\{y \in E\} \geq 1 - \nu$.

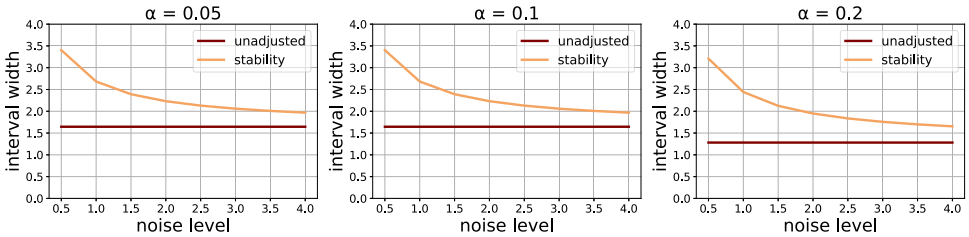

FIG. 3. *Normalized interval width implied by stability solution and unadjusted width for different noise levels b and error levels α. We fix ν = 0.05.*

Although in the main body of the paper we state the result when $\mathcal{P}_y = \mathcal{N}(\mu, \sigma^2)$ for simplicity, the Laplace noise addition strategy is valid for *arbitrary* $\mathcal{P}_y$; only the expression for η changes as a function of $\mathcal{P}_y$. In the Supplementary Material [54] we state a general version of Claim 5.

The proof of Claim 5 relies on showing that the selection to report is stable with respect to the oracle $\hat{S}(y'_E)$, and hence we can invoke Lemma 7.1.

We can see that, by choosing $\nu \to 0$, $b \to \infty$, we recover the nonselective confidence intervals, albeit at the cost of making the decision to report virtually independent of y. As we decrease b (and keep ν bounded away from zero), the decision to report becomes more reflective of the event report(y) and the inference level smoothly becomes more stringent. In Figure 3 we plot the normalized interval width, $z_{1-\frac{\alpha}{2}(1-\nu)e^{-\eta}}$, together with the unadjusted normalized width $z_{1-\frac{\alpha}{2}}$, for several different noise levels b and error levels α and $\nu = 0.05$.

As a concrete example, suppose that report$(y) = \{y > T\}$, for some threshold T. Then, as in Claim 2 and Claim 4, we can conclude that $\mathbf{1}\{\text{report}(y)\} = \mathbf{1}\{\text{report}(y + \xi)\}$ with probability $1 - \delta$ over the choice of ξ as long as $b \leq \frac{|y-T|}{\log(1/(2\delta))}$.

8. The design of stable selection algorithms. We discuss tools for designing stable algorithms and present an application of these tools to variable selection in linear regression. All proofs can be found in Appendix A of the Supplementary Material [54]. We begin with an overview of the basic properties of stability, which are key to efficient design of stable selections.

8.1. *Properties of stability.* Stability satisfies two key algorithmic properties: *closure under post-processing* and *composition*. We provide precise definitions of the two shortly. The reason why these properties enable efficient stability designs is that many selection rules can be written as post-processing and composition of simple computations, such as linear functions of the data or finding maxima of a sequence. As long as we know how to stabilize the necessary simple computations, closure under post-processing and composition provide rules for computing the overall stability parameter of the whole algorithm efficiently.

8.1.1. *Post-processing.* First, stability is *closed under post-processing*: if $\mathcal{A} : \mathbb{R}^n \to \mathcal{S}$ is (η, τ, ν)-stable, then for any (possibly randomized) map $\mathcal{B} : \mathcal{S} \to \mathcal{G}$, the composition $\mathcal{B} \circ \mathcal{A}$ is also (η, τ, ν)-stable. While the proof of this fact is a straightforward consequence of the definition of stability, the implications are significant. Suppose for the moment that the statistician is given a stable version of the LASSO algorithm, and denote its solution by $\hat{\theta}_{\text{LASSO}}$. Since $\hat{\theta}_{\text{LASSO}}$ is stable, then so is $\hat{M} = \{j \in [d] : \hat{\theta}_{\text{LASSO},j} \neq 0\}$. In fact, the statistician need not necessarily choose the model corresponding *exactly* to the support of $\hat{\theta}_{\text{LASSO}}$; for example, they could choose $\hat{M} = \{j \in [d] : |\hat{\theta}_{\text{LASSO},j}| \geq \epsilon\}$, for some constant threshold ϵ, or they could pick $d_{\text{sel}} \leq d$ entries with the maximum absolute value. More generally, any model

chosen solely as a function of $\hat{\theta}_{\text{LASSO}}$ inherits the same stability parameters as $\hat{\theta}_{\text{LASSO}}$. And, according to Corollary 6.1, the same PoSI constant suffices to correct the confidence intervals resulting from any such model.

Algorithm 1 Adaptive composition

input: data $y \in \mathbb{R}^n$, sequence of algorithms $\mathcal{A}_t : \mathcal{S}_1 \times \cdots \times \mathcal{S}_{t-1} \times \mathbb{R}^n \to \mathcal{S}_t$, $t \in [k]$
output: $(a_1, \ldots, a_k) \in \mathcal{S}_1 \times \cdots \times \mathcal{S}_k$
for $t = 1, 2, \ldots, k$ **do**
  Compute $a_t = \mathcal{A}_t(a_1, \ldots, a_{t-1}, y) \in \mathcal{S}_t$
end
Return $(a_1, \ldots, a_k)$

8.1.2. *Composition*. The second important property is *composition*. In Algorithm 1, we define adaptive composition. Adaptive composition consists of k sequential rounds in which the analyst observes the outcomes of all previous computations and selects the next computation *adaptively*—as a function of the previous evaluations. The adaptive composition property bounds the stability parameters of Algorithm 1 in terms of the stability parameters of $\mathcal{A}_t$. In its simplest form, it says that Algorithm 1 is $(k\eta, 0, 0)$-stable if for all $t \in [k]$, $\mathcal{A}_t(a_1, \ldots, a_{t-1}, \cdot)$ is $(\eta, 0, 0)$-stable for all fixed $a_1, \ldots, a_{t-1}$. For example, for some selection algorithms such as forward stepwise, it is clear to see how they can be represented using adaptive composition. In forward stepwise, $\mathcal{A}_t$ outputs an index $i_t \in [d]$, which corresponds to the variable i that minimizes the squared error resulting from adding i to the current pool of selected features; $i_t = \mathcal{A}_t(i_1, \ldots, i_{t-1}, y)$. It suffices to prove that any given step of forward stepwise selection is stable, in order to infer that the overall algorithm is stable as well.

Our proofs will only require adaptive composition for algorithms with $\nu = 0$; such results follow from classical theory on differential privacy. More advanced (and more conservative) adaptive composition theorems that allow $\nu > 0$ can be found in the context of typical stability [4]. In the Supplementary Material [54], we state the composition results we will need in our proofs.

A simpler kind of composition is nonadaptive composition. Here, the algorithms $\mathcal{A}_t$ have no dependence on the past computations. Nonadaptive composition can capture a protocol that involves running multiple selection methods and choosing a final selection target as an arbitrary function of all the outputs. As we state formally in the Supplementary Material, the resulting stability parameters simply add up. This is a rather appealing property of stability, as it suggests that the statistician only needs to keep track of the stability parameters of each selection algorithm they run, in order to derive valid selective confidence intervals. A similar combination of the results of different selection methods was considered by Markovic and Taylor [34].

8.2. *Model selection algorithms*: *Examples*. We now consider several algorithms for variable selection in linear regression through the lens of stability. While many of the principles presented in this section can be adapted to different distributional assumptions, for the sake of clarity and interpretability we assume that $y \sim \mathcal{N}(\mu, \sigma^2 I)$, where σ^2 is *unknown* but we have access to an estimate $\hat{\sigma}^2 \sim \sigma^2 \frac{\chi_r^2}{r}$, independent of y. This is the setup studied by Berk et al. [9]. More generally, we only need to know the decay of the tail of the distribution of y in order to enforce stability. In Appendix C of the Supplementary Material [54], we extend the algorithms in this section to outcome vectors with a known bound on their Orlicz norm, for any Orlicz function. This includes cases such as general sub-Gaussian and subexponential outcomes.

8.2.1. *Model selection via the LASSO.* We begin by considering the canonical example of the LASSO estimator [52]. The LASSO estimate is the solution to the usual least-squares problem with an additional ℓ_1-constraint on the regression coefficients:

$$\hat{\theta}_{\text{LASSO}} \in \underset{\theta \in \mathbb{R}^d}{\arg\min} \frac{1}{2}\|y - X\theta\|_2^2 \quad \text{s.t. } \|\theta\|_1 \leq C_1, \tag{4}$$

where $C_1 > 0$ is a tuning parameter. This problem is sometimes referred to as the LASSO in constrained/bound form, to contrast it with the LASSO in penalized form: $\hat{\theta}^{\lambda}_{\text{LASSO}} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1$, where $\lambda > 0$ is now the tuning parameter. These two problems are equivalent in a sense: given X and y, for any $C_1 > 0$ there exists a corresponding $\lambda > 0$ such that $\hat{\theta}_{\text{LASSO}} = \hat{\theta}^{\lambda}_{\text{LASSO}}$. In our analysis, we focus on the formulation (4). It is worth pointing out that our selective inference tools do not directly extend to penalized LASSO since, for a fixed penalty λ, the corresponding constraint C_1 depends on the data, which is random. Extending our approach to handle inference after solving the penalized problem is an important direction for future work.

The LASSO objective induces sparse solutions, and a common way of declaring that a feature is relevant is to check for a corresponding nonzero entry in the LASSO solution vector. That is, the model "selected" by the LASSO is

$$\hat{M} = \{j \in [d] : \hat{\theta}_{\text{LASSO},j} \neq 0\}.$$

Model selection via the LASSO was first analyzed in selective inference by Lee et al. [30]. While this work provides exact confidence intervals, it has been observed that the intervals (which do not make use of randomization) have infinite expected length [27]. Subsequent work has improved upon these often large confidence intervals by choosing a better event to condition on [33], or by applying randomization [26, 36, 37, 49–51].

We now formulate a stable version of the LASSO algorithm. It is inspired by the differentially private LASSO algorithm of Talwar et al. [46], although the noise variables are calibrated somewhat differently due to different modeling assumptions. We use e_i to denote the ith standard basis vector in $\mathbb{R}^d$, and $\{\pm e_i\}_{i=1}^d$ to denote the set of $2d$ standard basis vectors, multiplied by 1 and -1. We also let $\|X\|_{2,\infty} := \max_{i \in [d]} \|X_i\|_2$.

In essence, Algorithm 2 is a randomized version of the Frank-Wolfe algorithm [22].

We argue that $\hat{\theta}_{\text{LASSO}}$ is stable. The proof is based on a composition argument: namely, we can view $\hat{\theta}_{\text{LASSO}}$ as the result of a composition of k subroutines, each given by one optimization step which produces θ_t. The stability of each subroutine is proved by extending an argument related to the "report noisy max" mechanism from differential privacy [18].

Algorithm 2 Stable LASSO algorithm

input: design matrix $X \in \mathbb{R}^{n \times d}$, outcome vector $y \in \mathbb{R}^n$, variance estimate $\hat{\sigma}^2 \sim \sigma^2 \frac{\chi^2_r}{r}$, ℓ_1-constraint C_1, number of steps k, parameters $\delta \in (0, 1)$, $\eta > 0$
output: LASSO solution $\hat{\theta}_{\text{LASSO}} \in \mathbb{R}^d$
Initialize $\theta_1 = 0$
for $t = 1, 2, \ldots, k$ **do**
  $\forall \phi \in C_1 \cdot \{\pm e_i\}_{i=1}^d$, sample $\xi_{t,\phi} \overset{\text{i.i.d.}}{\sim} \text{Lap}(\frac{4t_{r,1-\delta/(2d)} C_1 \|X\|_{2,\infty}}{\eta n})$
  $\forall \phi \in C_1 \cdot \{\pm e_i\}_{i=1}^d$, let $\alpha_\phi = -\frac{2}{n\hat{\sigma}} \phi^\top X^\top (y - X\theta_t) + \xi_{t,\phi}$
  Set $\phi_t = \arg\min_{\phi \in C_1 \cdot \{\pm e_i\}_{i=1}^d} \alpha_\phi$
  Set $\theta_{t+1} = (1 - \Delta_t)\theta_t + \Delta_t \phi_t$, where $\Delta_t = \frac{2}{t+1}$
end
Return $\hat{\theta}_{\text{LASSO}} = \theta_{k+1}$

Algorithm 3 Stable marginal screening algorithm

input: design matrix $X \in \mathbb{R}^{n\times d}$, outcome vector $y \in \mathbb{R}^n$, variance estimate $\hat{\sigma}^2 \sim \sigma^2 \frac{\chi_r^2}{r}$, model size k, parameters $\delta \in (0,1)$, $\eta > 0$
output: $\hat{M} = \{i_1, \dots, i_k\}$
Compute $(c_1, \dots, c_d) = \frac{1}{n\hat{\sigma}} X^\top y \in \mathbb{R}^d$
$\text{res}_1 = [d]$
for $t = 1, 2, \dots, k$ **do**
  $\forall i \in \text{res}_i$, sample $\xi_{t,i} \overset{\text{i.i.d.}}{\sim} \text{Lap}(\frac{2t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n\eta})$
  $i_t = \arg\max_{i\in\text{res}_t} |c_i + \xi_{t,i}|$
  $\text{res}_{t+1} = \text{res}_t \setminus i_t$
end
Return $\hat{M} = \{i_1, \dots, i_k\}$

PROPOSITION 8.1 (LASSO stability). *Algorithm 2 is both*:

(a) $(\frac{1}{2}k\eta^2 + \sqrt{2k\log(1/\delta)}\eta, \delta, \delta)$-*stable*, *and*
(b) $(k\eta, 0, \delta)$-*stable*.

We state two rates because there exist parameter regimes where either rate leads to tighter confidence intervals than the other (the first rate being tighter when η is small).

By the post-processing property, Proposition 8.1 implies stability of any model $\hat{M}$ obtained as a function of $\hat{\theta}_{\text{LASSO}}$, such as the model corresponding to its nonzero entries.

Notice that the noise level in Algorithm 2 is an explicit function of η. This allows the statistician to understand the loss in utility—that is, how much worse $\hat{\theta}_{\text{LASSO}}$ is relative to an exact LASSO solution—due to randomization. In fact, building on work by Jaggi [24] and Talwar et al. [46], we can upper bound the excess risk resulting from randomization.

PROPOSITION 8.2 (LASSO utility). *Algorithm 2 run for* $k = \lceil \frac{n\|X\|_\infty^2 C_1\eta}{\hat{\sigma}\|X\|_{2,\infty}} \rceil$ *steps has*

$$\frac{1}{n}\mathbb{E}\big[\|y - X\hat{\theta}_{\text{LASSO}}\|_2^2 \mid y\big] - \min_{\theta:\|\theta\|_1\le C_1} \frac{1}{n}\|y - X\theta\|_2^2 = \tilde{O}\left(\frac{C_1\|X\|_{2,\infty}\log(d)t_{r,1-\delta/(2d)}\sigma}{n\eta}\right).$$

8.2.2. *Model selection via marginal screening*. One of the most commonly used model selection methods involves picking a constant number of the features with the largest absolute inner product with the outcome y [20, 23]. That is, one selects features i corresponding to the top k values of $|X_i^\top y|$, for a prespecified parameter k. This strategy is known as *marginal screening*. It was analyzed in the context of selective inference by Lee and Taylor [31].

In Algorithm 3, we state a stable version of marginal screening. Notice that the randomization scheme is similar to that of the stable LASSO method. Indeed, the high-level idea behind the proof of stability of Algorithm 3 is similar to that of Algorithm 2. As before, we let $\|X\|_{2,\infty}$ denote the $L_{2,\infty}$ norm of X.

PROPOSITION 8.3 (Marginal screening stability). *Algorithm 3 is both*:

(a) $(\frac{1}{2}k\eta^2 + \sqrt{2k\log(1/\delta)}\eta, \delta, \delta)$-*stable*, *and*
(b) $(k\eta, 0, \delta)$-*stable*.

As for the LASSO, we aim to quantify the loss in utility due to randomization. Given that the goal of marginal screening is to detect the largest k values $|c_i| = |X_i^\top y|$, a reasonable notion of utility loss is the difference between the values c_i corresponding to the variables in $\hat{M}$, and the actual largest values of c_i.

PROPOSITION 8.4 (Marginal screening utility). *Let m_i denote the index of the ith largest value c_j in absolute value, so that $(|c_{m_1}|, \dots, |c_{m_d}|)$ is the decreasing order statistic of $\{|c_i|\}_{i=1}^d$. Then, for any $\delta' \in (0,1)$, Algorithm 3 satisfies*

$$\mathbb{P}\left\{\max_{j\in[k]} |c_{m_j}| - |c_{i_j}| \le \frac{4t_{r,1-\delta/(2d)}\log(dk/\delta')\|X\|_{2,\infty}}{n\eta} | y\right\} \ge 1-\delta'.$$

9. Experimental results. We evaluate our selective intervals for the LASSO and marginal screening and compare our solution with data splitting.

For a fixed sample size n we vary the number of features d. We consider two different data-generating processes for the design matrix: one in which the rows of X are drawn independently from an equicorrelated multivariate Gaussian distribution with pairwise correlation 0.5, and the second one in which all entries of X are drawn as independent Bernoulli random variables with parameter 0.1. In the former case, X is normalized to have columns of unit norm. Given a signal parameter $\rho > 0$ and a sparsity parameter $s \in (0,1)$, we sample

$$\beta_i = \begin{cases} \text{Exp}(\rho), & i \in \{1, \dots, sd\}, \\ 0, & i \in \{sd+1, \dots, d\}, \end{cases}$$

and generate the outcome as $y = X\beta + \epsilon$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), i \in [n]$. In Appendix B of the Supplementary Material [54] we provide additional experiments when the errors are drawn from a heavier-tailed, Laplace distribution. We fix the target miscoverage level to be $\alpha = 0.1$. In all experiments we vary $\eta \in \{1,2,3,4,5,6,7,8,9,10\}$. For the comparison with data splitting, we use the splitting fraction derived in Section 5.1. Further experimental details are given in the Supplementary Material.

9.1. *Gaussian design.* We first state the results for the Gaussian design case.

In Figure 4 we compare the false discovery rate (FDR) of the stable LASSO algorithm and the LASSO algorithm with data splitting. In all plots $n = 50$ is fixed and we

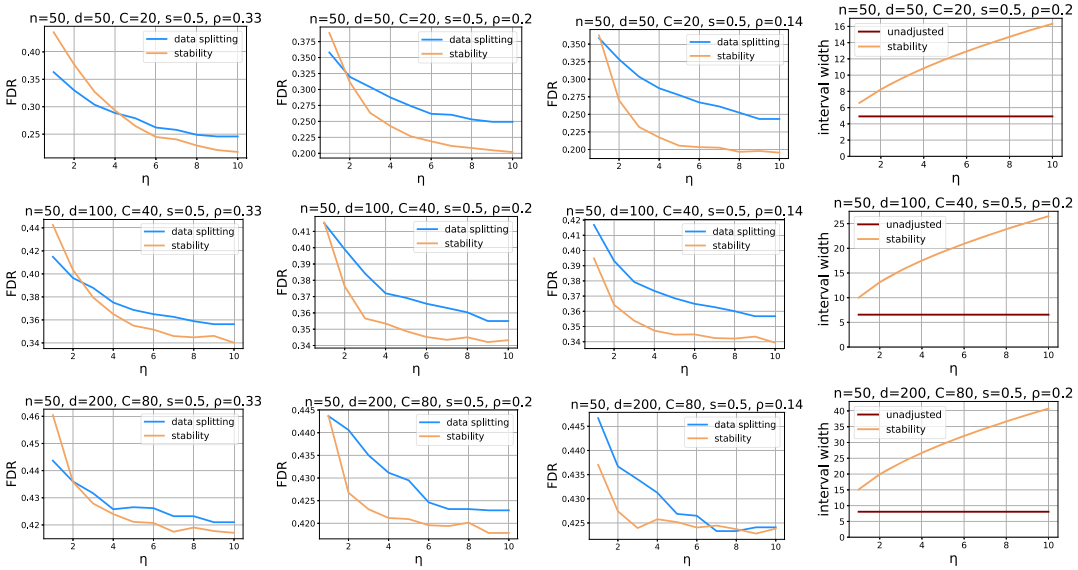

FIG. 4. *Comparison of FDR after stable LASSO and LASSO with data splitting, with varying dimension and signal strength, in the Gaussian design case. We also plot the average interval width (at $\rho = 0.2$ only, however the width varies minimally with ρ) and the average unadjusted width.*

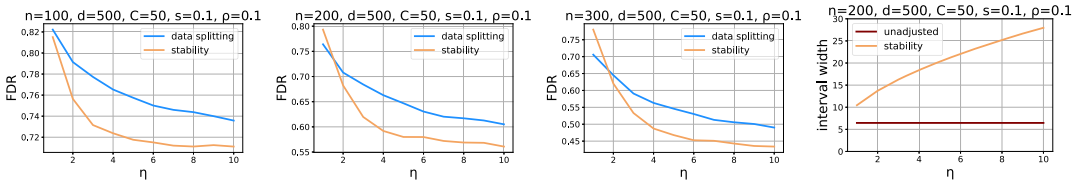

FIG. 5. *Comparison of FDR after stable LASSO and LASSO with data splitting, with varying sample size, in the Gaussian design case. We also plot the average interval width at* $n = 200$ *and the average unadjusted width.*

vary $d \in \{50, 100, 200\}$. As we increase d, we also increase the size of the constraint set $C_1 \in \{20, 40, 80\}$ to allow more selections. We consider signal levels $\rho \in \{0.33, 0.2, 0.14\}$, corresponding to an expected value of the nonnull β_i lying in $\{3, 5, 7\}$, and we fix $s = 0.5$.

We observe that stability generally outperforms data splitting as η grows, equivalently when the splitting fraction $f(\eta)$ grows, as well as when the signal strength grows. In Figure 4 we additionally plot the average width of stable intervals against the average width of naive, unadjusted intervals. Note that the intervals obtained via data splitting have essentially the same width (and are hence not plotted), based on how $f(\eta)$ is chosen. We only plot interval width for $\rho = 0.2$ since the width varies minimally for different values of ρ. For completeness we include all plots of interval width in Appendix B of the Supplementary Material [54].

In Figure 5 we compare the stable LASSO algorithm and the LASSO with data splitting in a sparse high-dimensional setting with $d = 500$, $s = 0.1$, and we vary the sample size $n \in \{100, 200, 300\}$. We fix $\rho = 0.1$. We observe that stability consistently outperforms data splitting for large enough η and this gap grows with n. In addition, we plot the average interval width implied by stability against the average unadjusted interval width at $n = 200$ (again we do not plot the interval width given by data splitting for the same reason as in Figure 4). We include the plots of all interval widths in the Supplementary Material.

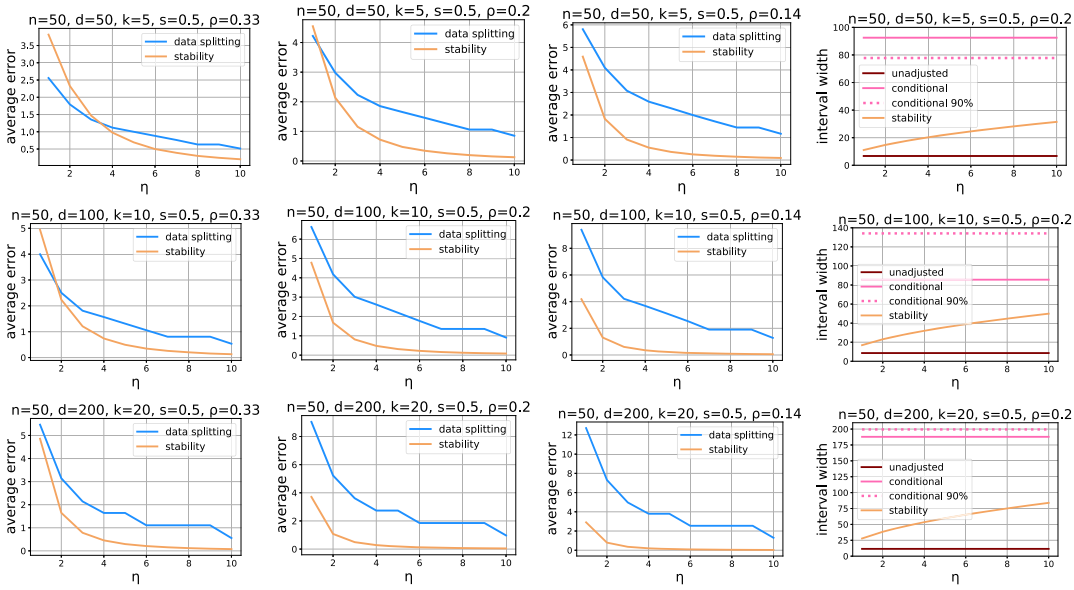

FIG. 6. *Comparison of average error after stable marginal screening and marginal screening with data splitting, with varying dimension and signal strength, in the Gaussian design case. In addition, we plot the average interval width (at* $\rho = 0.2$ *only, however the width varies minimally with* ρ*), together with the average unadjusted width and the width obtained via the conditional correction of Lee and Taylor* [31]*. We also plot the* 90% *quantile of the conditional width because it varies greatly across realizations.*

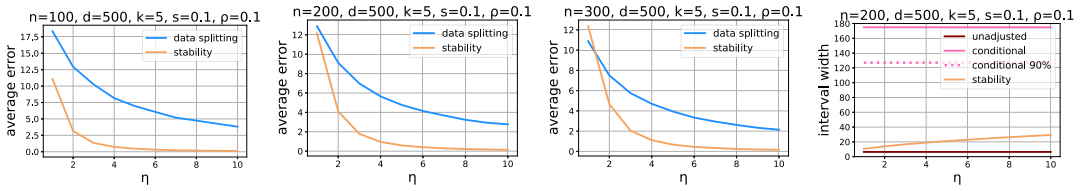

FIG. 7. *Comparison of average error after stable marginal screening and marginal screening with data splitting, with varying sample size, in the Gaussian design case. We also plot the average interval width at* $n = 200$*, together with the average unadjusted width and the width implied by the conditional approach of Lee and Taylor* [31].

In Figure 6 we compare the average error of stable marginal screening and marginal screening with data splitting. Since marginal screening explicitly aims to maximize the values $|X_i^\top y|$ for selected variables X_i, we quantify the error as $\frac{1}{k}\sum_{t=1}^{k}(|X_{i_t^*}^\top y| - |X_{i_t}^\top y|)$, where i_t is the estimated index of the tth largest absolute inner product (based on a subsample in the case of data splitting, or based on a randomized sample in the case of stability), and i_t^* is the true index of the tth largest absolute inner product in the data set. We vary the parameters as in the LASSO comparison in Figure 4, only instead of varying C_1 we vary $k \in \{5, 10, 20\}$. We also plot the average interval width with stability, together with the unadjusted interval width and the average width obtained via the conditional method of Lee and Taylor [31] with no randomization. For the conditional method, since the intervals are sometimes orders of magnitude larger than the average width, we also plot the 90% quantile of interval width. We see that stability typically outperforms data splitting in terms of the average error, and this benefit is more pronounced for larger η and signal strength. In terms of interval width, we observe that stability leads to significantly smaller intervals than the conditional approach. We only plot interval width when $\rho = 0.2$, and defer the remaining plots to the Supplementary Material.

In Figure 7 we consider a setting analogous to that of Figure 5, and we analogously vary the sample size n. We again see that stability generally dominates data splitting. Moreover, the

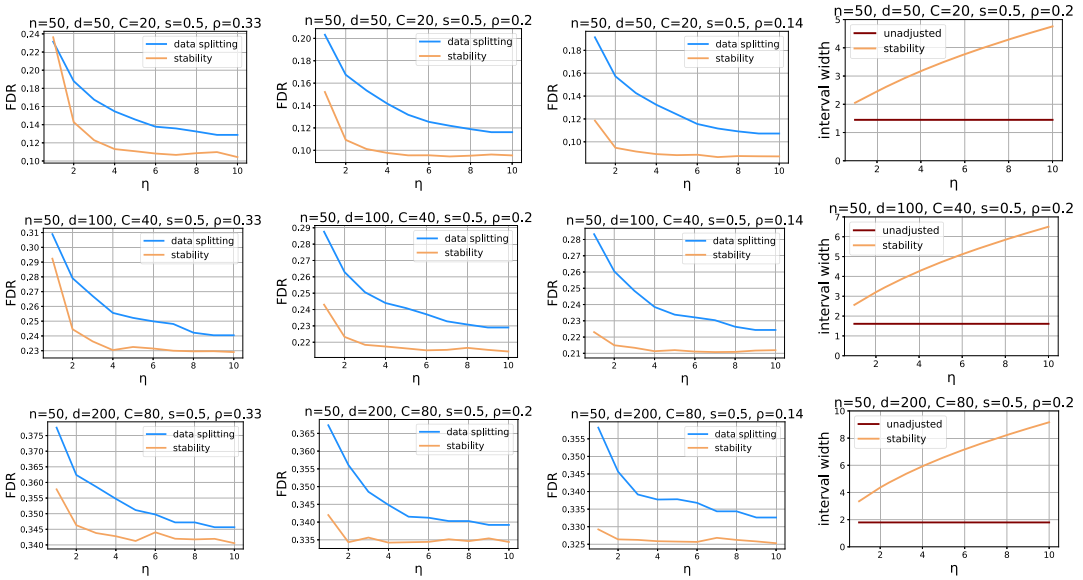

FIG. 8. *Comparison of FDR after stable LASSO and LASSO with data splitting, with varying dimension and signal strength, in the Bernoulli design case. We also plot the average interval width (at* $\rho = 0.2$ *only, however the width varies minimally with* ρ*) and the average unadjusted width.*

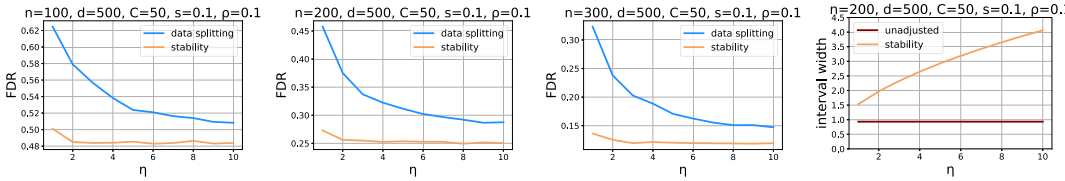

FIG. 9. *Comparison of FDR after stable LASSO and LASSO with data splitting, with varying sample size, in the Bernoulli design case. We also plot the average interval width at* $n = 200$ *and the average unadjusted width.*

gap between the intervals obtained via stability and those of Lee and Taylor [31] is even more pronounced than in Figure 6. We provide the plots of all interval widths in the Supplementary Material.

9.2. *Bernoulli design.* Now we consider the Bernoulli design case. The motivation for considering a sparse Bernoulli design lies in the fact that certain directions in the column space of X are captured by only a few samples, hence missing out on them—as is possible with data splitting—can significantly affect the quality of selection.

In Figure 8 and Figure 9 we provide comparisons analogous to those of Figure 4 and Figure 5, using the same parameter configurations. We observe a larger gap between data splitting and stability than in the Gaussian design case, and observe the same trends: as η and the signal strength grow, the performance gap increases. As before, we defer the remaining plots of interval widths to Appendix B of the Supplementary Material [54].

In Figure 10 and Figure 11 we provide comparisons analogous to those of Figure 6 and Figure 7, using the same parameter configurations. We observe a larger gap between data splitting and stability both than in the Gaussian design case, as well as in the LASSO experiments using the Bernoulli design. In addition, we observe an even more pronounced gap

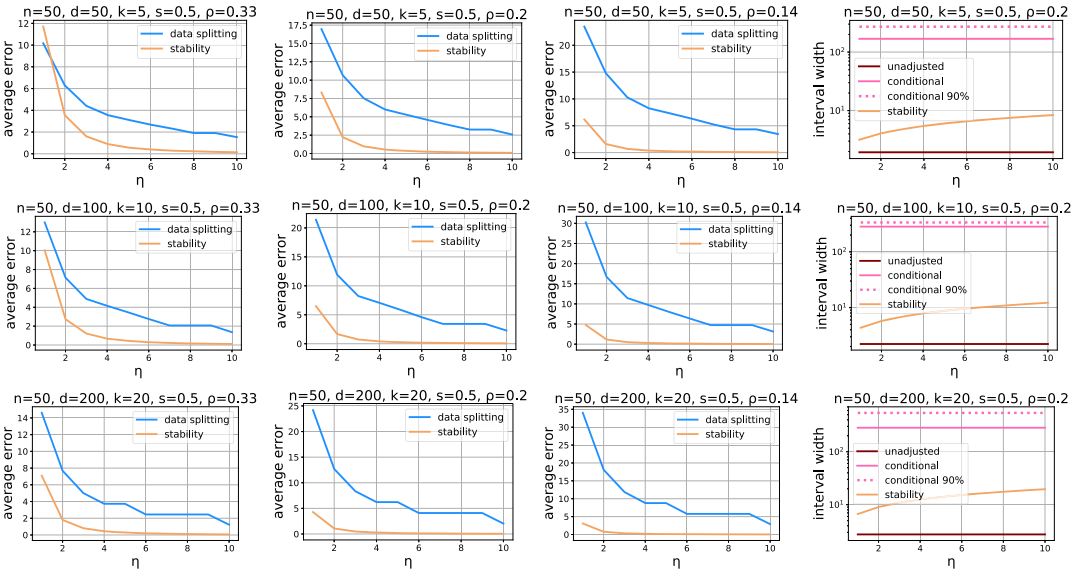

FIG. 10. *Comparison of average error after stable marginal screening and marginal screening with data splitting, with varying dimension and signal strength, in the Bernoulli design case. In addition, we plot the average interval width (at* $\rho = 0.2$ *only, however the width varies minimally with* ρ*), together with the average unadjusted width and the width obtained via the conditional correction of Lee and Taylor* [31]*. We also plot the* 90% *quantile of the conditional width because it varies greatly across realizations. Since the conditional widths are of a higher order of magnitude, the scale on the* y*-axis in the widths plots is logarithmic.*

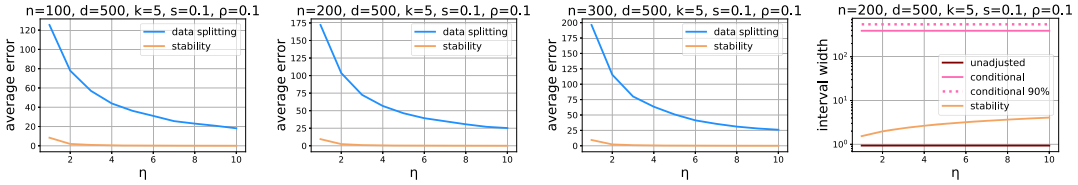

FIG. 11. *Comparison of average error after stable marginal screening and marginal screening with data splitting, with varying sample size, in the Bernoulli design case. We also plot the average interval width at* $n = 200$*, together with the average unadjusted width and the width implied by the conditional approach of Lee and Taylor* [31]*. Since the conditional widths are of a higher order of magnitude, the scale on the* y*-axis in the widths plot is logarithmic.*

between stable confidence interval widths and widths of intervals obtained via a conditional correction [31]. For this reason, the y-axis in the widths plots is logarithmic. We defer the remaining plots of interval widths to the Supplementary Material.

10. Discussion. Building on concepts from algorithmic stability, as originally developed for applications in differential privacy, we have provided general theory for designing post-selection confidence intervals when the selection procedure is stable. The stability principle is broadly applicable, ranging from inference on the winning effect to model selection in linear regression. In particular, stability is applicable even when data splitting is not, such as when there are dependencies between observations.

Performing inference after a stable selection is simple: it merely requires discounting the type I error based on the level of stability. Moreover, stability comes with several practically appealing properties, namely robustness to post-processing and composition. Thus, for example, the statistician can run various selection methods, and essentially only needs to keep track of the stability parameters of each in order to obtain valid confidence intervals for the final target, which could combine the results of all the selections in an arbitrary way.

There are numerous other potential applications of algorithmic stability to the problem of post-selection inference that would be worthwhile to explore. For example, it would be valuable to understand bootstrapping [39] from the perspective of stability, due to its conceptual relations to the "privacy amplification by subsampling" principle in differential privacy, which argues that privacy is amplified when run on a random subsample of the entire data set [6, 25]. More broadly, selective mechanisms have been long analyzed in the context of differential privacy [14, 19, 32, 45, 48], and we believe that some of these developments could be imported to selective inference via stability.

Acknowledgments. We are grateful to Vitaly Feldman, Will Fithian, Moritz Hardt, Arun Kumar Kuchibhotla, and Adam Sealfon for many helpful discussions and feedback which has lead to improvements of this work. In particular, we thank Will Fithian for pointing out the advantages of the oracle definition of stability.

Funding. This work was supported by the Army Research Office (ARO) under contract W911NF-17-1-0304 as part of the collaboration between US DOD, UK MOD and UK Engineering and Physical Research Council (EPSRC) under the Multidisciplinary University Research Initiative (MURI).

SUPPLEMENTARY MATERIAL

Supplement to "Post-selection inference via algorithmic stability" (DOI: 10.1214/23-AOS2303SUPP; .pdf). The supplement consists of three sections, containing proofs, deferred experimental results and simulation details, and stable algorithms for outcome vectors with bounded Orlicz norm.

REFERENCES

[1] ANDREWS, I., KITAGAWA, T. and MCCLOSKEY, A. (2019). Inference on winners. National Bureau of Economic Research.

[2] BACHOC, F., PREINERSTORFER, D. and STEINBERGER, L. (2020). Uniformly valid confidence intervals post-model-selection. *Ann. Statist.* **48** 440–463. MR4065169 https://doi.org/10.1214/19-AOS1815

[3] BARBER, R. F., CANDÈS, E. J., RAMDAS, A. and TIBSHIRANI, R. J. (2021). Predictive inference with the jackknife+. *Ann. Statist.* **49** 486–507. MR4206687 https://doi.org/10.1214/20-AOS1965

[4] BASSILY, R. and FREUND, Y. (2016). Typical stability. Preprint. Available at arXiv:1604.03336.

[5] BASSILY, R., NISSIM, K., SMITH, A., STEINKE, T., STEMMER, U. and ULLMAN, J. (2016). Algorithmic stability for adaptive data analysis. In *STOC'*16*—Proceedings of the* 48*th Annual ACM SIGACT Symposium on Theory of Computing* 1046–1059. ACM, New York. MR3536635

[6] BEIMEL, A., KASIVISWANATHAN, S. P. and NISSIM, K. (2010). Bounds on the sample complexity for private learning and private data release. In *Theory of Cryptography. Lecture Notes in Computer Science* **5978** 437–454. Springer, Berlin. MR2673387 https://doi.org/10.1007/978-3-642-11799-2_26

[7] BENJAMINI, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biom. J.* **52** 708–721. MR2758547 https://doi.org/10.1002/bimj.200900299

[8] BENJAMINI, Y., HECHTLINGER, Y. and STARK, P. B. (2019). Confidence intervals for selected parameters. Preprint. Available at arXiv:1906.00505.

[9] BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

[10] BI, N., MARKOVIC, J., XIA, L. and TAYLOR, J. (2020). Inferactive data analysis. *Scand. J. Stat.* **47** 212–249. MR4075236 https://doi.org/10.1111/sjos.12425

[11] BOUSQUET, O. and ELISSEEFF, A. (2002). Stability and generalization. *J. Mach. Learn. Res.* **2** 499–526. MR1929416 https://doi.org/10.1162/153244302760200704

[12] BUJA, A., BROWN, L., BERK, R., GEORGE, E., PITKIN, E., TRASKIN, M., ZHANG, K. and ZHAO, L. (2019). Models as approximations I: Consequences illustrated with linear regression. *Statist. Sci.* **34** 523–544. MR4048582 https://doi.org/10.1214/18-STS693

[13] CUMMINGS, R., LIGETT, K., NISSIM, K., ROTH, A. and WU, Z. S. (2016). Adaptive learning with robust generalization guarantees. In *Conference on Learning Theory* (*COLT*) 772–814.

[14] DURFEE, D. and ROGERS, R. M. (2019). Practical differentially private top-k selection with pay-what-you-get composition. In *Advances in Neural Information Processing Systems* (*NeurIPS*) 3532–3542.

[15] DWORK, C., FELDMAN, V., HARDT, M., PITASSI, T. and REINGOLD, O. (2015). Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems* (*NIPS*) 2350–2358.

[16] DWORK, C., FELDMAN, V., HARDT, M., PITASSI, T., REINGOLD, O. and ROTH, A. (2015). Preserving statistical validity in adaptive data analysis [extended abstract]. In *STOC'*15*—Proceedings of the* 2015 *ACM Symposium on Theory of Computing* 117–126. ACM, New York. MR3388189

[17] DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography. Lecture Notes in Computer Science* **3876** 265–284. Springer, Berlin. MR2241676 https://doi.org/10.1007/11681878_14

[18] DWORK, C. and ROTH, A. (2013). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9** 211–487. MR3254020 https://doi.org/10.1561/0400000042

[19] DWORK, C., SU, W. and ZHANG, L. (2015). Private false discovery rate control. Available at arXiv:1511.03803.

[20] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

[21] FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Available at arXiv:1410.2597.

[22] FRANK, M. and WOLFE, P. (1956). An algorithm for quadratic programming. *Nav. Res. Logist. Q.* **3** 95–110. MR0089102 https://doi.org/10.1002/nav.3800030109

[23] GUYON, I. and ELISSEEFF, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3** 1157–1182.

[24] JAGGI, M. (2013). Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the* 30*th International Conference on Machine Learning* 427–435.

[25] KASIVISWANATHAN, S. P., LEE, H. K., NISSIM, K., RASKHODNIKOVA, S. and SMITH, A. (2011). What can we learn privately? *SIAM J. Comput.* **40** 793–826. MR2823508 https://doi.org/10.1137/090756090

[26] KIVARANOVIC, D. and LEEB, H. (2020). A (tight) upper bound for the length of confidence intervals with conditional coverage. Available at arXiv:2007.12448.

[27] KIVARANOVIC, D. and LEEB, H. (2021). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *J. Amer. Statist. Assoc.* **116** 845–857. MR4270029 https://doi.org/10.1080/01621459.2020.1732989

[28] KUCHIBHOTLA, A. K., BROWN, L. D., BUJA, A. and CAI, J. (2019). All of linear regression. Preprint. Available at arXiv:1910.06386.

[29] KUCHIBHOTLA, A. K., BROWN, L. D., BUJA, A., CAI, J., GEORGE, E. I. and ZHAO, L. H. (2020). Valid post-selection inference in model-free linear regression. *Ann. Statist.* **48** 2953–2981. MR4152630 https://doi.org/10.1214/19-AOS1917

[30] LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

[31] LEE, J. D. and TAYLOR, J. E. (2014). Exact post model selection inference for marginal screening. In *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*. **1** 136–144. https://doi.org/10.5555/2968826.2968842

[32] LEI, J., CHAREST, A.-S., SLAVKOVIC, A., SMITH, A. and FIENBERG, S. (2018). Differentially private model selection with penalized and constrained likelihood. *J. Roy. Statist. Soc. Ser. A* **181** 609–633. MR3807500 https://doi.org/10.1111/rssa.12324

[33] LIU, K., MARKOVIC, J. and TIBSHIRANI, R. (2018). More powerful post-selection inference, with application to the Lasso. Preprint. Available at arXiv:1801.09037.

[34] MARKOVIC, J. and TAYLOR, J. (2016). Bootstrap inference after using multiple queries for model selection. Preprint. Available at arXiv:1612.07811.

[35] KUCHIBHOTLA, A. K., RINALDO, A. and WASSERMAN, L. (2020). Berry-Esseen bounds for projection parameters and partial correlations with increasing dimension. Preprint. Available at arXiv:2007.09751.

[36] PANIGRAHI, S., MARKOVIC, J. and TAYLOR, J. (2017). An MCMC-free approach to post-selective inference. Available at arXiv:1703.06154.

[37] PANIGRAHI, S. and TAYLOR, J. (2022). Approximate selective inference via maximum likelihood. *J. Amer. Statist. Assoc.* 1–11.

[38] RASINES, D. G. and YOUNG, G. A. (2023). Splitting strategies for post-selection inference. *Biometrika* **110** 597–614. MR4627773 https://doi.org/10.1093/biomet/asac070

[39] RINALDO, A., WASSERMAN, L. and G'SELL, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.* **47** 3438–3469. MR4025748 https://doi.org/10.1214/18-AOS1784

[40] ROGERS, R., ROTH, A., SMITH, A. and THAKKAR, O. (2016). Max-information, differential privacy, and post-selection hypothesis testing. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 487–494. IEEE Computer Soc., Los Alamitos, CA. MR3631011 https://doi.org/10.1109/FOCS.2016.59

[41] ROSENTHAL, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull.* **86** 638.

[42] RUSSO, D. and ZOU, J. (2016). Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics* 1232–1240.

[43] SCHEFFÉ, H. (1999). *The Analysis of Variance*. *Wiley Classics Library*. Wiley, New York. MR1673563

[44] STEINBERGER, L. and LEEB, H. (2023). Conditional predictive inference for stable algorithms. *Ann. Statist.* **51** 290–311. MR4564857 https://doi.org/10.1214/22-aos2250

[45] STEINKE, T. and ULLMAN, J. (2017). Tight lower bounds for differentially private selection. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2017 552–563. IEEE Computer Soc., Los Alamitos, CA. MR3734260 https://doi.org/10.1109/FOCS.2017.57

[46] TALWAR, K., THAKURTA, A. G. and ZHANG, L. (2015). Nearly optimal private LASSO. In *Advances in Neural Information Processing Systems* (*NIPS*) 3025–3033.

[47] TAYLOR, J. and TIBSHIRANI, R. J. (2015). Statistical learning and selective inference. *Proc. Natl. Acad. Sci. USA* **112** 7629–7634. MR3371123 https://doi.org/10.1073/pnas.1507583112

[48] THAKURTA, A. G. and SMITH, A. (2013). Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory* (*COLT*) 819–850.

[49] TIAN HARRIS, X., PANIGRAHI, S., MARKOVIC, J., BI, N. and TAYLOR, J. (2016). Selective sampling after solving a convex problem. Preprint. Available at arXiv:1609.05609.

[50] TIAN, X., BI, N. and TAYLOR, J. (2017). MAGIC: a general, powerful and tractable method for selective inference. Preprint. Available at arXiv:1607.02630.

[51] TIAN, X. and TAYLOR, J. (2018). Selective inference with a randomized response. *Ann. Statist.* **46** 679–710. MR3782381 https://doi.org/10.1214/17-AOS1564

[52] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[53] TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* **111** 600–620. MR3538689 https://doi.org/10.1080/01621459.2015.1108848

[54] ZRNIC, T. and JORDAN, M. I (2023). Supplement to "Post-selection inference via algorithmic stability." https://doi.org/10.1214/23-AOS2303SUPP