Online Control of the False Coverage Rate and False Sign Rate

Asaf Weinstein [*1] **Aaditya Ramdas** [*2]

Abstract

The reproducibility debate has caused a renewed interest in changing how one reports uncertainty, from p-value for testing a null hypothesis to a confidence interval (CI) for the corresponding parameter. When CIs for multiple selected parameters are being reported, the analog of the false discovery rate (FDR) is the false coverage rate (FCR), which is the expected ratio of number of reported CIs failing to cover their respective parameters to the total number of reported CIs. Here, we consider the general problem of FCR control in the online setting, where one encounters an infinite sequence of fixed unknown parameters ordered by time. We propose a novel solution to the problem which only requires the scientist to be able to construct marginal CIs. As special cases, our framework yields algorithms for online FDR control and online sign-classification procedures that control the false sign rate (FSR). All of our methodology applies equally well to prediction intervals, having particular implications for selective conformal inference.

1. Introduction

While statisticians are trained to be aware of multiple testing issues, temporal multiplicity—when decisions need to be made one at a time, rather than in a batch—is often easy to miss. Let us examine the following simplified situation. Consider a team at a pharmaceutical company who test a new drug every week. In week t, a new drug is under inspection, and to assess its treatment effect θ_t, the team conducts a new randomized clinical trial with new participants. Suppose that the data, such as the normalized empirical difference in means between the treatment and control groups, can be summarized by the observation $X_t \sim N(\theta_t, 1)$, independent of all the previous observations $\{X_i\}_{i<t}$.

Now consider the following selection rule: if $X_t < 3$, then the analysts simply ignore drug t. If $X_t > 3$, then the team reports the two-sided 99% marginal interval $X_t \pm 2.58$ for θ_t to the management, who may then decide to run a much larger second phase clinical trial since the CI does not contain 0. This may initially seem like an innocuous situation: each drug is different and has a different treatment effect θ_t, the data X_t is always fresh and independent of the data for other drugs, and the decision for whether or not to construct the CI for θ_t (and the interval itself, if constructed) is dependent only on X_t and independent of all other X_i.

Nonetheless, the combination of multiplicity and selection is a cause for concern already in the offline setting, as was insightfully pointed out by Benjamini & Yekutieli (2005). In the online case, when there is an infinite sequence of parameters, it is even easier to construct an example where ignoring selection has undesirable consequences. Indeed, consider the special case where $\theta_t = 0$ for all t, in other words, every tested drug is equivalent to a placebo. We could equivalently assume that $\theta_t \in \{-0.1, +0.1\}$, i.e., each drug is either slightly worse or better than the placebo. Then, every single CI that is reported to the management is incorrect, since $X_t \pm 2.58$ does not contain zero when $X_t > 3$. Because (an infinite number of) selections will occur with probability one, the fraction of non-covering CIs—the false coverage proportion, FCP—will equal one. Of course, the second phase of the trial may correct these errors, but at the cost of time, money and faith in the analyst.

In the example above, the problem is that the CI is reported only if X_t is large. Indeed, $X_t \pm 2.58$ includes θ_t with probability 0.99 unconditionally or even conditionally on which CIs were reported up to time $t-1$, but not including time t itself; we refer to such an interval as a "marginal" CI (and define it formally in Section 3). By contrast, a *conditional* CI would satisfy

$$\Pr\{\theta_t \in \text{CI} \mid \theta_t \text{ is selected}\} \geq 1 - \alpha; \qquad (1)$$

constructing a conditional CI at each step will rectify the situation and it controls the FCR at any time T (see Section B in supplement). This conditional solution is conceptually simple but has two main drawbacks. First, it is impossible to ensure that a reported CI never includes zero. This is demon-

[*]Equal contribution [1]School of Computer Science and Engineering, Hebrew University of Jerusalem [2]Departments of Statistics & Data Science, and Machine Learning, Carnegie Mellon University. Correspondence to: Asaf Weinstein <asaf.weinstein@mail.huji.ac.il>, Aaditya Ramdas <aramdas@cmu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

strated in Section 5 and in Section B.1 of the supplement. Second, constructing a conditional CI may be a prohibitive task depending on how complicated the selection event and the (unconditional) distribution of X_i are.

In this paper, we will propose a solution for online FCR control that is very different from the aforementioned conditional approach. Informally, in order to achieve FCR control at level α, instead of constructing $(1-\alpha)$ conditional CIs, we construct $(1-\alpha_i)$ marginal CIs for some $\alpha_i < \alpha$. The algorithm to set the α_i is inspired by recent advances in the online false discovery rate (FDR) control literature, specifically the LORD algorithm by Javanmard & Montanari (2018) and its uniform improvement by Ramdas et al. (2017). The new online FCR procedure works in much more generality than the simple example described above, for example when θ_i are multi-dimensional, the data is not necessarily Gaussian, the selection events are complicated, and so on—cases in which constructing a conditional CI may be substantially harder if at all possible.

Even more importantly, by constructing marginal instead of conditional CIs, we leave open the possibility to use the candidate CI itself as a criterion for selection (the decision whether or not to report the CI). For example, the rule may entail constructing the candidate marginal CI only if it does not include values of opposite signs. Thus, returning to the motivating example, this would allow the team of statisticians to ensure that each reported CI is conclusive about the *direction* of the treatment effect, while the FCR is controlled. With such situations in mind, we can instantiate our procedure with sign-determining CIs, resulting in an online adaptation of the ideas of Weinstein & Yekutieli (2014). Every such sign-determining CI procedure corresponds to an online sign-classification procedure that controls the false sign rate (FSR). As a special case, we show that for some recent online testing procedures, supplementing rejections based on two-sided p-values with directional decisions suffices to control the FSR.

The following point is critical: while a procedure controlling FCR can be used to derive a procedure controlling FSR or FDR, the converse is not true. Indeed, there are a slew of procedures for offline and online FDR control, but most of these have no analog for the CI problem. We are aware of only one general procedure for offline FCR control (Benjamini & Yekutieli, 2005), and in this paper we propose the first and only existing procedure for online FCR control.

The rest of this paper is organized as follows. Section 2 sets up the problem formally and introduces necessary notation. A new online procedure that adjusts marginal confidence intervals, is presented in Section 3. In Section 4 we show how our marginal CI procedure can be used to solve a general online localization problem, and study the special case of the online sign-classification problem. Simulation results that also compare the marginal approach and the conditional approach, are reported in Section 5. We conclude with a brief discussion in Section 6, where we mention how all of our results also hold for prediction intervals for unseen responses. Most proofs appear in the supplement.

2. Problem Setup

Let $\theta_1, \theta_2, \ldots$ be a fixed sequence of fixed unknown parameters, where the domain Θ_i of θ_i is arbitrary, but common examples include $\mathbb{R}$ or $\mathbb{R}^d$. Let 2^{Θ_i} denote the set of all measurable subsets of Θ_i, in other words it is any acceptable confidence set for θ_i. In our setup, at each time step i, we observe an independent observation (or summary statistic) $X_i \in \mathcal{X}_i$, where the distribution of X_i depends on θ_i (and possibly other parameters). For example, when $\Theta_i = \mathbb{R}$ we may have $X_i \sim N(\theta_i, 1)$. Let $\mathcal{S}_i : \mathcal{X}_i \to \{0,1\}$ denote the selection rule, that indicates whether or not the user wishes to report a confidence set for θ_i. Also, denote by $S_i := \mathcal{S}_i(X_i)$ the indicator for selection, where $S_i = 1$ means that the user will report a confidence set for θ_i. Let the filtration (increasing sequence of sigma-fieldss) formed by the selection decisions be denoted by

$$\mathcal{F}^i := \sigma(S_1, S_2, \ldots, S_i).$$

Next, let $\mathcal{I}_i : \mathcal{X}_i \times [0,1] \to 2^{\Theta_i}$ be the rule for constructing the confidence set for θ_i, the second argument allowing to take as an input a "confidence level". We denote $I_i := \mathcal{I}_i(X_i, \alpha_i)$. Thus, $I_i = \mathcal{I}_i(X_i, \alpha_i)$ may be a marginal or a conditional $(1-\alpha_i)$ confidence set for θ_i as discussed later, but in general it is no more than a mapping from $\mathcal{X}_i \times [0,1]$ as described above. For simplicity, in the rest of the paper we refer to I_i as a confidence *interval* (CI) like it would usually be if $\Theta_i = \mathbb{R}$, but with the understanding that everything discussed in this paper applies to the more general case of arbitrary confidence sets. The above rules are all required to be *predictable*, meaning

$$\mathcal{S}_i, \mathcal{I}_i, \alpha_i \text{ are } \mathcal{F}^{i-1}\text{-measurable},$$

and we write $\mathcal{S}_i, \mathcal{I}_i, \alpha_i \in \mathcal{F}^{i-1}$. Of course, the instantiated random variables $I_i = \mathcal{I}_i(X_i, \alpha_i), S_i = \mathcal{S}_i(X_i)$ both depend on X_i. However, the *rules* $\mathcal{S}_i, \mathcal{I}_i, \alpha_i$ must be $\mathcal{F}^{i-1}$-measurable, hence specified before observing X_i.

Using the terminology above, we now define an online selective-CI procedure. In the rest of the paper we omit the term "selective", but this is done only for the sake of readability. Thus, an *online CI protocol* proceeds as follows:

1. At time i, first commit to $\alpha_i, \mathcal{S}_i, \mathcal{I}_i \in \mathcal{F}^{i-1}$.
2. Then, observe X_i, and decide whether or not θ_i is selected for coverage by setting $S_i = \mathcal{S}_i(X_i)$.
3. Report $I_i = \mathcal{I}_i(X_i, \alpha_i)$ if $S_i = 1$. Then, increment i, and go back to step 1.

We remark that we could have also defined the filtration as $\widetilde{\mathcal{F}}^i := \sigma(X_1, \ldots, X_i)$ and all algorithms/proofs would still go through; we use $\widetilde{\mathcal{F}}^i$ because it may be more difficult to construct a useful CI using the conditional distribution $X_i|\widetilde{\mathcal{F}}^i$ than using $X_i|\mathcal{F}^i$ if the X_is are dependent (there is less residual randomness in the former); under independence of X_is, this would not be a concern. Note that $\mathcal{S}_i$ can depend on $\mathcal{I}_i$ because both are predictable, and hence S_i can depend on I_i—for example, we can select a parameter for coverage based on whether or not I_i looks "favorable", a point to which we will return in later sections.

To evaluate the errors made by an online CI protocol, let the unobservable false coverage indicator be denoted $V_i := S_i \mathbb{1}_{\theta_i \notin I_i}$. The false coverage proportion up to time T is defined as the ratio of the number of miscovering reported intervals to the total number of reported intervals:

$$\text{FCP}(T) := \frac{\sum_{i \le T} V_i}{\sum_{i \le T} S_i},$$

where $0/0 = 0$ per standard convention (i.e., if no intervals are constructed, then the false coverage proportion is trivially zero). The false coverage rate (FCR) and the modified FCR are defined, respectively, as

$$\text{FCR}(T) = \mathbb{E}\left[\frac{\sum_{i \le T} V_i}{\sum_{i \le T} S_i}\right], \quad \text{mFCR}(T) = \frac{\mathbb{E}\left[\sum_{i \le T} V_i\right]}{\mathbb{E}\left[\sum_{i \le T} S_i\right]}.$$

The main objective of this paper is to develop and compare algorithms for specifying $\mathcal{I}_i$ and α_i such that FCR or mFCR control is guaranteed at any time, that is,

$$\text{FCR}(T) \le \alpha \;\; \forall T \in \mathbb{N}, \qquad \text{or} \qquad \text{mFCR}(T) \le \alpha \;\; \forall T \in \mathbb{N}.$$

We will later observe that procedures that control the FCR can also be used to control other error metrics of interest.

3. The LORD-CI algorithm

By a *marginal* (unconditional) CI rule $\mathcal{I}_i$, we mean that

$$\Pr\left\{\theta_i \notin \mathcal{I}_i(X_i, a) \;\middle|\; \mathcal{F}^{i-1}\right\} \le a \text{ for any } a \in [0,1], \quad (2)$$

where the probability is taken only over randomness in X_i (conditional on $\mathcal{F}^{i-1}$), because the rule $\mathcal{I}_i$ is predictable. Importantly, we use the term 'marginal' because a conditional CI would have further conditioned on $S_i = 1$.

In what follows, an *algorithm* is a sequence of mappings from past selection decisions to confidence levels, i.e. from $(S_1, \ldots, S_{i-1})$ to α_i. By definition, such an α_i is $\mathcal{F}^{i-1}$-measurable, hence a procedure that constructs a marginal confidence interval at level $(1 - \alpha_i)$ whenever $S_i = 1$, is an online CI protocol. We will refer to such a procedure as a *marginal online CI protocol/procedure*. A trivial marginal online CI protocol can be obtained by taking any *fixed* sequence of α_i such that the series $\sum_{i=1}^{\infty} \alpha_i \le \alpha$; this procedure is referred to as alpha-spending by Foster & Stine (2008) in the context of online hypothesis testing, and it controls the familywise error rate, which in our context is the probability of even a single miscoverage event. Naturally, this is a much more stringent notion of error, and the resulting selected CIs will be excessively wide. The question we will address below is the following: is there a nontrivial algorithm to set the α_i so that FCR is controlled?

3.1. mFCR control for arbitrary selection rules

Our first result identifies a sufficient condition on an algorithm to imply mFCR control. For this, associate any algorithm with an estimate of the false coverage proportion,

$$\widehat{\text{FCP}}(T) := \frac{\sum_{i \le T} \alpha_i}{(\sum_{i \le T} S_i) \vee 1}.$$

We may then define the following procedure.

Definition 1 (LORD-CI procedure)**.** A LORD-CI procedure is any online protocol that constructs marginal $(1 - \alpha_i)$ confidence intervals, where $\alpha_i \in \mathcal{F}^{i-1}$ are defined in a predictable fashion to maintain the invariant

$$\forall T \in \mathbb{N}, \quad \widehat{\text{FCP}}(T) \le \alpha. \quad (3)$$

Any LORD-CI procedure has the following guarantee.

Theorem 1. *Given an arbitrary sequence of selection rules made by the user, any LORD-CI procedure enjoys* $\text{mFCR}(T) \le \alpha$, $\forall T \in \mathbb{N}$.

Proof. By definition of a false coverage event, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i \le T} V_i\right] &= \mathbb{E}\left[\sum_{i \le T} S_i I_{\theta_i \notin I_i}\right] \overset{(a)}{\le} \sum_{i \le T} \mathbb{E}\left[I_{\theta_i \notin I_i}\right] \\
&= \sum_{i \le T} \mathbb{E}\left[\mathbb{E}\left[I_{\theta_i \notin I_i} \;\middle|\; \mathcal{F}^{i-1}\right]\right] \overset{(b)}{\le} \sum_{i \le T} \mathbb{E}\left[\alpha_i\right] \\
&= \mathbb{E}\left[\sum_{i \le T} \alpha_i\right] \overset{(c)}{\le} \alpha \mathbb{E}\left[\sum_{i \le T} S_i\right],
\end{aligned}$$

where inequality (a) holds because $S_i \le 1$, inequality (b) by the definition (2) of a $(1 - \alpha_i)$ marginal CI, and inequality (c) by the invariance (3). Rearranging the first and last expression yields the desired result. □

If one really insisted on requiring FCR control as opposed to mFCR control, we provide a guarantee for a subclass of "monotone" selection rules, as introduced below.

3.2. Monotonicity of algorithms, CIs and selection rules

The symbol $\succeq$ is used to compare vectors coordinatewise, so $(v) \succeq (w)$ means that $v_i \geq w_i$ for all i.

An online FCR algorithm is called *monotone* if for any two vectors $(s_1, \ldots, s_{i-1}) \succeq (\widetilde{s}_1, \ldots, \widetilde{s}_{i-1})$, we have $\alpha_i(s_1, \ldots, s_{i-1}) \geq \alpha_i(\widetilde{s}_1, \ldots, \widetilde{s}_{i-1})$. Equivalently, an online FCR algorithm is monotone if

$$\alpha_i \geq \widetilde{\alpha}_i \quad \text{whenever } (S_1, \ldots, S_{i-1}) \succeq (\widetilde{S}_1, \ldots, \widetilde{S}_{i-1}),$$

where $\widetilde{\alpha}_i$ is the level produced by the online FCR algorithm, when presented with the history of selection decisions $(\widetilde{S}_1, \ldots, \widetilde{S}_{i-1})$. A CI rule $\mathcal{I}$ is called *monotone* if

$$\mathcal{I}(x, a_2) \subseteq \mathcal{I}(x, a_1) \text{ for all } a_1 < a_2 \text{ and } x \in \mathcal{X}.$$

Monotonicity is satisfied for most natural (even non-equivariant) CI constructions, and thus we do not view this as a restriction. Irrespective of whether the online FCR algorithm and CI rule are monotone, we say that a selection rule $\mathcal{S}_i$ is *monotone* if

$$S_i \geq \widetilde{S}_i \text{ whenever } (S_1, \ldots, S_{i-1}) \succeq (\widetilde{S}_1, \ldots, \widetilde{S}_{i-1}), \quad (4)$$

where, as before, $\widetilde{S}_i$ is used to denote the selection decision at time i, for the same observation X_i, but for a different history $(\widetilde{S}_1, \ldots, \widetilde{S}_{i-1})$.

As a simple special case, if each rule $\mathcal{S}_i$ is independent of $\mathcal{F}^{i-1}$, then such a selection rule is trivially monotone, even if the underlying online FCR algorithm is not. In other words, if the final decision $S_i = \mathcal{S}_i(X_i)$ is based only on X_i and on none of the past decisions, then such a rule is monotone. For example, setting $S_i = \mathbb{1}_{X_i > 3}$ for every i is a trivial monotone selection rule.

3.3. FCR control for monotone selection rules

We can offer the following guarantee for the nontrivial class of monotone selection rules.

Theorem 2. *Given an arbitrary sequence of monotone selection rules (4) chosen by the user, any LORD-CI procedure that maintains the invariant (3) has the guarantee* $\forall T \in \mathbb{N}, \mathrm{FCR}(T) \leq \alpha$.

Proof. By definition of $\mathrm{FCR}(T)$, we have

$$\mathrm{FCR}(T) = \mathbb{E}\left[\frac{\sum_{i \leq T} V_i}{\sum_{j \leq T} S_j}\right] = \sum_{i \leq T} \mathbb{E}\left[\frac{S_i \mathbb{1}_{\theta_i \notin I_i}}{\sum_{j \leq T} S_j}\right] \leq \sum_{i \leq T} \mathbb{E}\left[\frac{\alpha_i}{\sum_{j \leq T} S_j}\right], \quad (5)$$

where the inequality follows by Lemma 1 below. Thus, we see that

$$\mathrm{FCR}(T) \leq \mathbb{E}\left[\frac{\sum_{i \leq T} \alpha_i}{\sum_{j \leq T} S_j}\right] \leq \alpha,$$

where the last inequality holds due to invariant (3). □

The critical step in the above argument is the invocation of the following central lemma (proved in Appendix A).

Lemma 1. *Given an arbitrary sequence of monotone selection rules, we have*

$$\mathbb{E}\left[\frac{S_i \mathbb{1}_{\theta_i \notin I_i}}{\sum_{j \leq T} S_j}\right] \leq \mathbb{E}\left[\frac{\alpha_i}{\sum_{j \leq T} S_j}\right].$$

Intuitively, the statement of the above lemma is obvious if the expectation could be taken separately in the numerator, as if it were independent of the denominator, because $\mathbb{E}[S_i \mathbb{1}_{\theta_i \notin I_i}] \leq \mathbb{E}[\mathbb{1}_{\theta_i \notin I_i}] \leq \alpha_i$ by construction (2). The proof, included in the supplement, demonstrates that monotonicity allows us to formally perform such a step. Lemma 1 is a generalization of lemmas that have been proved in the context of online FDR control by Javanmard & Montanari (2018); Ramdas et al. (2017; 2018); Tian & Ramdas (2019), since the selection event $\{S_i = 1\}$ may or may not be associated with the miscoverage event $\{\theta_i \notin I_i\}$, but in online FDR control, the rejection event $\{R_i = 1\} \equiv \{P_i \leq \alpha_i\}$ is obviously directly related to the false discovery event $\{P_i \leq \alpha_i, i \in \mathcal{H}_0\}$. Indeed, we will later see that online FCR control captures online FDR control as a special case.

3.4. An explicit monotone online FCR algorithm

An uncountable number of procedures satisfy the conditions of Theorems 3 and 2. To provide one concrete example of a monotone update rule that satisfies (3), we adapt the LORD++ online FDR algorithm (Javanmard & Montanari, 2018; Ramdas et al., 2017) in order to set the sequence $\{\alpha_i\}$. LORD++ was originally designed to maintain an invariant resembling (3) in the context of multiple hypothesis testing, i.e., when $S_i = 1$ amounts to rejection of the i-th null hypothesis. In the absence of p-values, our algorithm instead substitutes rejection events by arbitrary selection events $(S_i = 1)$. We refer to the aforementioned adaptation of LORD++ to the context of CIs as *the LORD-CI algorithm/procedure*. Whenever we refer to LORD-CI, we mean the online protocol as given in the algorithm box below.

In Algorithm 1, $\{\gamma_i\}_{i=1}^{\infty}$ is a deterministic nonincreasing sequence of positive constants summing to one, that is specified in advance; $0 \leq W_0 \leq \alpha$ is a prespecified constant; $\{\mathcal{S}_i\}$ is a sequence of arbitrary predictable selection rules; and $\{\mathcal{I}_i\}$ is a sequence of marginal CIs, that is, each $\mathcal{I}_i$ has the property (2). On implementing Protocol 1, set $\gamma_i = 0$ whenever $i \leq 0$ (this happens for $j = 1$). It is easy to verify

that α_i (see line 7 in Protocol 1) is monotone, because it has an additional nonnegative term in the summation with every new selection. One may also verify that $\{\alpha_i\}$ satisfies the invariant (3), because $\sum_i \alpha_i$ is always less than $(\sum_i S_i)\alpha$.

Algorithm 1 A monotone instantiation of LORD-CI

Input : sequence $\{X_i\}$ observed sequentially; prespecified deterministic sequence $\{\gamma_i\}$ summing to one; constant $W_0 \in (0, \alpha)$; arbitrary selection rules $\{\mathcal{S}_i\}$; marginal CI rules $\{\mathcal{I}_i\}$.
Output : online FCR-adjusted selective CIs
$t \leftarrow 1$ // tracks time
for $j = 1, 2, \ldots$ **do**
  while $\mathcal{S}_i(X_i) = 0$ **do**
    $t \leftarrow t + 1$ // θ_t is not selected, increment time
  end
  $\tau_j \leftarrow t$ // time of the j-th selection
  $\alpha_t \leftarrow \gamma_t W_0 + (\alpha - W_0)\gamma_{t-\tau_1} + \alpha \sum_{\{k:\tau_k<t, \tau_k \neq \tau_1\}} \gamma_{t-\tau_k}$
  Report $I_t = \mathcal{I}_t(X_t, \alpha_t)$
  $t \leftarrow t + 1$
end

The slightly complicated form of α_i above really boils down to the definition of FCP or FCR. In fact, if their denominators were replaced by $1 + \sum_{i \le T} S_i$, and the same was done for $\widehat{\text{FCP}}$, then the above rule would simplify to $\alpha_t \leftarrow \alpha \sum_{\{k:\tau_k<t\}} \gamma_{t-\tau_k} \equiv \alpha \sum_{k=1}^{j-1} \gamma_{t-\tau_k}$. Intuitively, each selection 'earns' the analyst an extra error budget of α and this newly earned budget is spread across future times using the sequence $\{\gamma_i\}$. Thus, at time t, the current round's error budget α_t is determined by summing the separate amount that each past earning contributed towards time t.

4. Selections that depend on the candidate CIs

In the LORD-CI procedure, the sequence of the $\mathcal{S}_i$ and the sequence of the $\mathcal{I}_i$ are both arbitrary, and these may be specified independently of each other. In this section we demonstrate how tying the selection rule to the CI rule, by letting the (candidate) marginal CI *determine* whether θ_i is selected or not, can lead to many instantiations of the LORD-CI procedure that are of practical interest.

Informally, the idea is as follows. Suppose that we made some choice in advance for the marginal CI rule. Suppose also that we have in mind a criterion for what constitutes a "good" reported CI. For example, when $\Theta_i \equiv \mathbb{R}$, a "good" CI might be one that excludes zero. Then at each step i, upon observing X_i, we pretend that we were to construct $I_i = \mathcal{I}(X_i, \alpha_i)$ where α_i are set by the LORD-CI algorithm, but we only actually report it if it is "good". By design, then, we only report "good" intervals. Note that because $\mathcal{I}_i$ is predictable, choosing to report an interval only if it is "good" is a *predictable* selection rule. Therefore we may use LORD-CI to determine levels of coverage and immediately be guaranteed FCR control. This is formalized in Definition 2 below. We remark that these ideas have appeared in Weinstein & Yekutieli (2014), but their treatment is rather informal, and theirs is not an online procedure.

In what follows, whenever we use a CI rule, it is assumed to be monotone. Again, we do not view this as a restriction.

4.1. From coverage to localization

Suppose that for each i we have a collection $K_{i1}, \ldots, K_{iL_i} \in 2^{\Theta_i}$ of pre-specified disjoint subsets of Θ_i. Being able to say that $\theta_i \in K_{il}$ for exactly one $l \in \{1, \ldots, L_i\}$ qualifies as having "localized" the signal.[1] On observing X_i, we must either localize θ_i by specifying which of $1, \ldots, L_i$ it belongs to, or refrain from making any claim at all about θ_i. The natural notion of error for a given procedure is a *false localization rate* (FLR),

$$\text{FLR} := \mathbb{E}\left[\frac{\#\text{false localizations made}}{\#\text{localizations made}}\right].$$

As we will see below, the false localization rate generalizes the false discovery rate.

Definition 2 (LORD-CI for localization)**.** Let $\mathcal{I}_i : \mathcal{X}_i \times [0,1] \to 2^{\Theta_i}$ be an arbitrary pre-specified monotone marginal CI rule for $i = 1, 2, \ldots$, and define $\mathcal{S}_i$ as follows:

$$S_i = \begin{cases} 1, & \text{if } I_i = \mathcal{I}_i(X_i, \alpha_i) \text{ is a subset of exactly} \\ & \text{one of } K_{i1}, \ldots, K_{iL_i} \\ 0, & \text{otherwise} \end{cases}.$$

Then *LORD-CI for localization* is the online CI protocol that applies LORD-CI to the above selection rule, and when $S_i = 1$, it outputs the unique $j_i \in [L_i]$ with $I_i \subseteq K_{ij_i}$.

Theorem 3. *The LORD-CI for localization procedure (Definition 2) satisfies* $\text{FLR}(T) \le \alpha$ *for any* T.

Next we consider some special cases of localization and their implications; Table 1 summarizes the different notions introduced in the ensuing subsections.

4.2. Composite hypothesis testing with FDR control

Suppose that we have a sequence of composite null hypotheses to test, $H_0^i : \theta_i \in \Theta_{0i}$, $i = 1, 2, \ldots$, where $\Theta_{0i} \subseteq \Theta_i$. For any online testing procedure, let $R_i := \mathbb{1}(H_0^i \text{ is rejected})$ and define

$$\text{FDR}(T) := \mathbb{E}\left[\frac{\#\{i \le T : R_i = 1, \theta_i \in \Theta_{0i}\}}{\#\{i \le T : R_i = 1\}}\right],$$

[1] If the sets are not disjoint a-priori, one may either create a new set for the intersection, or generalize the definition of localization to allow for the reporting of multiple sets.

Localization	Error metric	Procedure
point null testing	FDR	Definition 1
composite null testing	FDR	Definition 3
sign-classification	FSR	Definition 4

Table 1. Summary of localization problems considered. The first column indicates the type of the localization problem, the second column is the specialized FLR, and the third column is the corresponding instantiation of LORD-CI.

which reduces to the usual definition of the FDR when Θ_i include a single value, i.e., when testing point null hypotheses. We can use the procedure of Definition 2 to devise an online *testing* protocol that controls the FDR.

Definition 3 (LORD-CI for composite testing)**.** Consider an arbitrary marginal CI rule $\mathcal{I}_i$ for θ_i. We reject the ith composite null hypothesis and set $R_i = 1$, if and only if

$$\mathcal{I}_i(X_i, \alpha_i) \cap \Theta_{0i} = \emptyset, \tag{6}$$

where α_i is set by the LORD-CI procedure using $S_i = R_i$.

The above procedure comes with the following guarantee.

Corollary 1. *The LORD-CI procedure for composite testing (Definition 3) enjoys* $\mathrm{FDR}(T) \leq \alpha$ *for any* T.

Before proceeding, we would like to point out a connection to existing online testing protocols. We can define a p-value for testing H_0^i by $P_i := \sup\{\alpha : \mathcal{I}_i(X_i, \alpha) \cap \Theta_{0i} \neq \emptyset\}$, where $\mathcal{I}_i$ is any monotone CI for θ_i. Indeed, if $\theta_i \in \Theta_{0i}$, then for any $t \in [0, 1]$ we have

$$\begin{aligned}\Pr_{\theta_i}\{P_i \leq t\} &\leq \Pr_{\theta_i}\{\mathcal{I}_i(X_i, t) \cap \Theta_{0i} = \emptyset\} \\ &\leq \Pr_{\theta_i}\{\theta_i \notin \mathcal{I}_i(X_i, t)\} \leq t.\end{aligned}$$

Therefore, we can apply an existing online FDR protocol using the above p-value. Note that while the computation of the p-value above might not be trivial, we are really only required at each step to check if $P_i \leq \alpha_i$, which is equivalent to rejecting when $\mathcal{I}_i(X_i, t) \cap \Theta_{0i} = \emptyset$. In fact, if we use the same CI rules and the same algorithm to set the α_i as in the CI procedure employed in Definition 3, we obtain exactly the composite testing procedure of Definition 3.

4.3. Sign-classification with FSR control

Sometimes we would like to ask about the *direction* of the effect rather than test a two-sided null hypothesis. As argued in e.g. Gelman et al. (2012) and Gelman & Tuerlinckx (2000), this often makes a more sensible question than asking whether a parameter is *equal* to zero. In fact, even statisticians that use a two-sided test of a point null hypothesis, tend to supplement—perhaps with a leap of faith—a rejection of the null with a claim about the sign of the parameter (Goeman et al., 2010, call this *post hoc* inference of the sign).

In Section 1, the management might be interested primarily in identifying which drugs have a positive/non-positive treatment effect (significantly better/worse than placebo). Throughout this subsection suppose that $\Theta_i \equiv \Theta \subseteq \mathbb{R}$, $\mathcal{X}_i \equiv \mathcal{X}$ and that $X_i \sim f(x_i; \theta_i)$ for some common likelihood function $f : \mathcal{X} \times \Theta \to [0, \infty)$, so that a common CI rule can be used at all times—note that a CI rule depends on the likelihood function only, not on the true value of θ_i—and we call this situation the "common likelihood" case.

When considering a sign-classification procedure, we will aim to control—in analogy to the FDR—the expected ratio of number of incorrect directional decisions to the total number of directional decisions made. Throughout this paper, to make a directional decision means to classify $\theta_i > 0$ (positive) or $\theta_i \leq 0$ (non-positive); because zero is included on one side, this can be considered a *weak* sign-classification (although the definition is not symmetric, zero can be just as well be appended to the positive side instead of the negative side). Hence, a sign-classification protocol is an online procedure that outputs

$$D_i = \begin{cases} 1, & \text{if } \theta_i \text{ classified as positive} \\ -1, & \text{if } \theta_i \text{ classified as non-positive} \\ 0, & \text{if no decision on the sign of } \theta_i \text{ is made} \end{cases}.$$

Borrowing a term from Stephens (2016), we define the *false sign rate* (FSR) as

$$\mathrm{FSR}(T) := \mathbb{E}\left[\frac{\#\{i \leq T : (\theta_i \leq 0, D_i = 1) \text{ or } (\theta_i > 0, D_i = -1)\}}{\#\{i \leq T : D_i = 1 \text{ or } D_i = 1\}}\right].$$

In the realistic situation where there are no parameters that equal zero exactly, this coincides with the definitions of Benjamini et al. (1993). A natural procedure to consider is applying any online FDR protocol to test the hypotheses that $\theta_i = 0$, and then classify each rejection according to the sign of an unbiased estimate of θ_i. So, for example, when $\theta_i = \mathbb{E}[X_i]$, a rejected null with $X_i > 0$ entails $D_i = 1$. We will see later that applying LORD++ to the usual two-sided p-values indeed works, however FSR control is not *automatically* guaranteed, i.e., this requires a proof.

We rely again on the procedure of Definition 2 to devise a sign-classification protocol that controls the FSR. Thus, suppose that we have an arbitrary (common) marginal CI rule $\mathcal{I}(\cdot, \cdot)$. Now specialize the prescription in Definition 2 by taking $L_i \equiv 2$, $K_{i1} \equiv (-\infty, 0], K_{i2} \equiv (0, \infty)$. In words, this is the LORD-CI procedure that reports $I_i = \mathcal{I}(X_i, \alpha_i)$ whenever it includes either only positive or only non-positive values. This instantiation of the procedure in Definition 2 is central enough to merit a separate definition.

Definition 4 (Sign-determining LORD-CI procedure)**.** Suppose that we are in the "common likelihood" case, and let

$\mathcal{I} : \mathcal{X} \times [0,1] \to 2^{\Theta}$ be any marginal confidence interval. Assume that for any parameter θ_i there is a corresponding "null" value $\theta_{0i} \in \Theta$. The *sign-determining LORD-CI* procedure associated with $\mathcal{I}$ is defined to be the LORD-CI procedure that utilizes the selection rules

$$S_i = \begin{cases} 1, & \text{if } \{\tau - \theta_{0i} : \tau \in \mathcal{I}(x, \alpha_i)\} \subseteq (0, \infty) \quad \text{or} \\ & \{\tau - \theta_{0i} : \tau \in \mathcal{I}(x, \alpha_i)\} \subseteq (-\infty, 0] \\ 0, & \text{otherwise} \end{cases}, \tag{7}$$

and constructs $I_i = \mathcal{I}(X_i, \alpha_i)$ if $S_i = 1$.

For simplicity, assume from now on that $\theta_{0i} \equiv 0$. In that case the sign-determining LORD-CI procedure constructs $I_i = \mathcal{I}(X_i, \alpha_i)$ if and only if this interval is *sign-determining*, meaning that it includes only positive or only nonpositive values.

Returning to the FSR problem, apply now the CI procedure from Definition 4 with an arbitrary choice of $\mathcal{I}$, and set

$$D_i = \begin{cases} 1, & \text{if } S_i = 1 \text{ and } I_i \subseteq (0, \infty) \\ -1, & \text{if } S_i = 1 \text{ and } I_i \subseteq (-\infty, 0] \\ 0, & \text{if } S_i = 0 \end{cases}. \tag{8}$$

Then we have the following result:

Corollary 2. *The sign-classification procedure given by* (8)*, enjoys* $\text{FSR}(T) \leq \alpha$ *for any* T.

Remark 1. It is easy to see that we could drop the assumption on monotonicity of the CI rules in this section, and still be guaranteed control of the respective modified error rates, for example mFDR in Subsection 4.2 and mFSR in Subsection 4.3 (which would now be implied by mFCR control for the corresponding CI procedure).

Notice that *any* marginal CI rule defines a sign-determining LORD-CI procedure. As such, the choice of the marginal CI determines both the power as a sign-classification procedure and the shape/length of the corresponding intervals. Ideally, the sign-determining LORD-CI procedure would construct a large number of CIs, and at the same time ensure that the lower endpoint for a positive interval is as far away from zero as possible (similarly, the upper endpoint for a nonpositive interval is as far away from zero as possible). As demonstrated with the two elementary choices of the marginal CI given below, these goals are conflicting in general (see also Benjamini et al., 1998). For the rest of this section assume $X_i \sim N(\theta_i, 1)$, though the constructions that follow can be extended beyond the normal case.

1. $\mathcal{I}$ is the usual two-sided CI, $\mathcal{I}(x, \alpha) = (x - z_{\alpha/2}, x + z_{\alpha/2})$. It can be verified that the sign-determining LORD-CI procedure with this choice for $\mathcal{I}$, selects exactly the set of parameters rejected by the LORD++ online FDR procedure using the two-sided p-values

   $$P_i = 2(1 - \Phi(|X_i|)).$$

   A reported CI has length $2z_{\alpha_i/2}$. It is worth mentioning that, on interpreting this as a sign-classification procedure through (8), corollary 2 would hold even if $D_i = -1$ is interpreted as declaring that $\theta_i < 0$ instead $\theta_i < 0$, because the interval is open.

2. $\mathcal{I}$ is the "one-sided" interval given by

   $$\mathcal{I}(x, \alpha) = \begin{cases} (x - z_\alpha, x + z_\alpha), & \text{if } 0 < |x| < z_\alpha \\ (0, x + z_\alpha), & \text{if } x > z_\alpha \\ (x - z_\alpha, 0], & \text{if } x < -z_\alpha \end{cases}.$$

   The sign-determining LORD-CI procedure now selects exactly the set of parameters rejected by the LORD++ online FDR procedure using the one-sided p-values

   $$P_i = 1 - \Phi(|X_i|),$$

   hence is much more powerful as a sign-classification procedure. However, the length of the constructed intervals is unbounded, since they take the form $(0, X_i + z_{\alpha_i})$ or $(X_i - z_{\alpha_i}, 0]$. More seriously, reported intervals must touch zero, thus failing to address our follow-up question on how big the effect is.

Instead of insisting on maximizing power or minimizing length, we can choose a marginal CI that is a compromise between the two choices of $\mathcal{I}$ presented above. For example, the MQC interval of Weinstein & Yekutieli (2014) determines the sign earlier than the two-sided interval but not as early as the "one-sided" interval. In turn, it leads to more power than LORD++ applied to two-sided p-values (interpreted as a sign-classification procedure), and at the same time it separates from zero for sufficiently large x. Figure 4 in Appendix C is taken from Weinstein & Yekutieli (2014) and shows the endpoints of this CI for $\alpha = 0.1$.

Section 5 demonstrates the potential gain in power due to using the MQC interval.

5. Simulation

We carry out experiments where online confidence intervals are constructed under different (predictable) selection schemes. Setting $\alpha = 0.1$, in each of $N = 10,000$ simulations we draw $m = 10,000$ parameters i.i.d. from a mixture

$$\theta_i = \begin{cases} 0.5\delta_{10^{-3}} + 0.5\delta_{-10^{-3}}, & \text{w.p. } 0.9 \\ 1 + W_i, & \text{w.p. } 0.1 \end{cases},$$

where $W_i \sim \text{Pois}(1)$. The mass at $\pm 10^{-3}$ represents the "null" component (essentially zero), while the "nonnulls"

are drawn so that large effects are rare. We then draw the observations $X_i \sim N(\theta_i, 1)$. In our implementation, the LORD-CI algorithm always uses the sequence of α_i of the LORD++ procedure (Ramdas et al., 2017) with $W_0 = \alpha/2$, $\gamma(j) = 0.0722\frac{\log(j\vee 2)}{je^{\sqrt{\log j}}}$, as in the experiments of Javanmard & Montanari (2018); Ramdas et al. (2017). For a conditional CI we used the construction from Weinstein et al. (2013, Section 2) obtained by inverting shortest acceptance regions.

In the first simulation a CI is constructed when $|X_i| > 3$. Figure 1 shows conditional CIs versus LORD-CI intervals for a single realization. The conditional CIs are considerably shorter than LORD-CI; in particular, the conditional CIs resemble the marginal two-sided 90% interval for large observations. LORD-CI appears conservative with FCP $= 0.043$ as compared to about 0.1 for conditional. Both the conditional and LORD-CI intervals may cross zero, as shown in the plot. in fact, as many as 53% of the conditional CIs cross zero, and 38% of LORD-CI intervals cross zero. Note that the lower endpoint of the CI is monotone non-decreasing for the conditional intervals, but not for LORD-CI. The conditional intervals seem preferable in this situation (and indeed this is a situation where it is easy to construct a conditional interval if so desired, but this should be thought of as rare).

The second simulation example illustrates a situation where we are interested first in detecting the sign of the parameters, and second in supplementing a directional decision with confidence bounds. Parameters are selected relying on the sign-determining LORD-CI procedure of Section 4, that is, θ_i is selected whenever the candidate LORD-CI interval excludes zero. We are free to choose the (monotone) marginal CI used with the LORD-CI algorithm: the symmetric two-sided marginal CI amounts to selecting according to the LORD++ testing protocol utilizing two-sided p-values, and supplementing each selection with a decision on the sign according to the sign of X_i. Instead, to increase power, we can use a marginal CI that itself has early sign determination, as explained in Section 4. We implemented both, first equipping the sign-determining LORD-CI procedure with the symmetric interval, and second with the MQC interval of Weinstein & Yekutieli (2014). The latter always selects at least as many parameters as the former. In both cases, the constructed CIs are (by design) compatible with the primary decision on the sign, i.e., none of the CIs includes values of opposite signs. As in the first simulation, we constructed also conditional CIs as a competitor to our method.

Table 2 displays averages over the N rounds as a summary of the simulation. Except for the FCR for LORD-CI intervals being excessively small, the results are in line with our expectations. Figure 2 displays adjusted LORD-CI marginal CIs along with conditional CIs for the selected parameters, for a single realization of the experiment. MQC resulted in a 13% increase in the number of parameters selected.

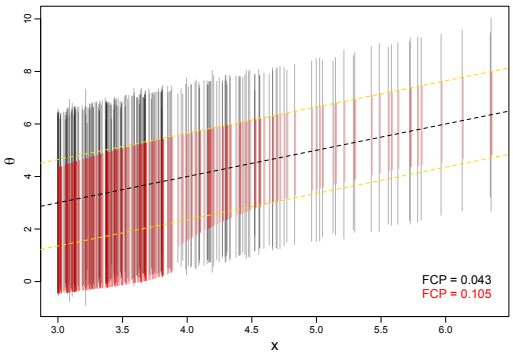

Figure 1. LORD-CI vs. conditional CIs, selection above a fixed threshold. The figure shows constructed intervals as vertical segments, conditional CIs in red and LORD-CI in black. Black dashed line is the identity line, and the yellow dashed lines represent the endpoints of the marginal (unadjusted) 90% interval.

	Symmetric		MQC	
	LORD-CI	cond.	LORD-CI	cond.
FCR	0.03	0.1	0.032	0.1
mFCR	0.031	0.1	0.032	0.1
$\mathbb{E}\sum_i S_i$	133.5	133.5	154.4	154.4
$\mathbb{E}\frac{\sum_i S_i \cdot A_i}{\sum_i S_i}$	1	0.533	1	0.527

Table 2. Summary for the second simulation example. In the headline, "Symmetric" and "MQC" indicate what marginal CI equips the LORD-CI procedure. In the last row, A_i is the indicator for the event that the reported CI is sign-determining.

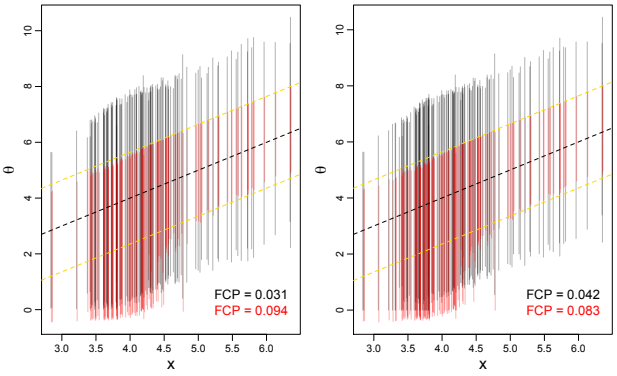

Figure 2. LORD-CI vs. conditional CIs, selection by the sign-determining LORD-CI procedure. The figure shows constructed intervals as vertical segments: conditional CIs in red, LORD-CI in black. Black dashed line is the identity line, and the yellow (diagonal, light) dashed lines represent the endpoints of the marginal (unadjusted) 90% interval. *Left*: selection utilizes the usual symmetric interval. *Right*: selection utilizes the MQC interval.

The conditional CIs are significantly shorter than LORD-CI for large observations. However, the conditional intervals may cross zero. In fact, more than half (51.2%) of the conditional CIs include both positive and negative values. As mentioned in the introduction, in many applications it would be desirable to report CIs that do not cross zero, and in that sense LORD-CI has here a clear advantage. In practice, one might be tempted to remove the conditional CIs that include zero; of course, with this strategy we lose FCR control—in our example, the FCP would increase to 0.2 if we kept only the intervals excluding zero. An elaborate discussion is included in Section B.1.

6. Discussion

In a multiplicity problem, a sequence of unknown parameters $\theta_i,\ i \leq m \in \mathbb{N} \cup \{\infty\}$ is considered simultaneously, and for each we observe data X_i that is informative about θ_i. The main task is usually to use the observations to localize "interesting" parameters in the sequence, while controlling in some specified sense the rate of erroneous localizations.

In realistic situations, it is often the case that: (i) the X_i arrive one by one, and we have to make decisions on the fly; (ii) we are concerned with more complicated problems than testing a sequence of point null hypotheses. For example, we might want to know which θ_i are bigger than 10 and which are smaller than -5, as opposed to simply testing whether $\theta_i > 0$; and (iii) there is a follow-up question to the primary one, e.g., if a drug can be declared to improve by at least 10%, we would also like to know also how large the improvement is at most.

We offer methodology to address all of the concerns in the list above. The basic tool is an online algorithm for adjusting the levels of *selective* marginal CIs so that we have control over the false coverage rate for any predictable selection rule. We show how to instantiate our online CI procedure to provide answers to items 2 and 3 in the list above: adapted to the example above, we would offer an online procedure that reports only CIs that are either contained in $(10, \infty)$ or contained in $(-\infty, -5)$, while ensuring that FCR $\leq \alpha$ at any point in time. The power of the online FCR procedure as a localization rule, and the shape/length of the reported CIs, is determined by the choice of the underlying marginal CI rule being used. The CI rules are allowed to be arbitrary in this paper, and we offered some insights on their interplay with FSR control. The methodology itself applies to any observation space $\mathcal{X}$ and any parameter space Θ.

We note that except for LORD++, we do not think that FCR analogs exist for other known online FDR procedures such as alpha-investing (Foster & Stine, 2008), generalized alpha-investing (Aharoni & Rosset, 2014), or the recent "adaptive" algorithms like SAFFRON (Ramdas et al., 2018) or ADDIS (Tian & Ramdas, 2019).

From a practical viewpoint, the procedure of Definition 2 may have important implications. While we did not mention this explicitly in Section 4.1, L_i and the sets K_{il} may more generally be *predictable* rather than fixed (the guarantees on FLR control will not change). This allows considerable flexibility because by choosing these adaptively, the statistician is able to change the *type* of decision made at each time step. For example, at $t = 1$ she may be interested in testing a null hypothesis, at $t = 2$ in classifying the sign of θ_t, and at $t = 3$ in constructing an interval that includes only values bigger than, say, 0.5. **Thus, one might say that the LORD-CI algorithm really controls the 'false decision rate', where the type of decision can itself be determined online.** This observation is particularly useful in applied large-scale experimentation where the aim of each experiment is usually different, and a joint error rate may be otherwise hard to define (and control).

We end this article with an important remark. Keeping with historical treatment of multiple testing problems, it felt natural to write this article in terms of parameters and CIs, but our entire methodology for FCR control goes through seamlessly for *prediction intervals* as well. For instance, consider a regression problem where we are given a training dataset $(X_j^{\text{tr}}, Y_j^{\text{tr}}) \sim P_X \times P_{Y|X}$ and a sequence of test points $X_i \sim P_X$ at which we may want to make predictions. We may treat the unknown Y_i in the same way as we treated the unknown θ_i in this paper. On observing X_i, we may decide whether we wish to report a prediction interval I_i for Y_i or not (and this decision S_i could be based on the previous selection decisions $\mathcal{F}^{i-1}$ and on I_i). As long as for each i, we can construct marginally valid intervals, i.e.,

$$\Pr\{Y_i \in \mathcal{I}_i(X_i, \alpha_i) \mid \mathcal{F}^{i-1}\} \leq \alpha_i,$$

then the FCR (or other variants) for the selected prediction intervals will be controlled exactly as in the case of selected confidence intervals.

The above insight has particularly important ramifications for *conformal prediction*, which is a way of constructing distribution-free marginal predictive intervals (Vovk et al., 2005; Shafer & Vovk, 2008). Indeed, it is also well known that distribution-free conditional inference is impossible (Vovk, 2012; Lei & Wasserman, 2014; Barber et al., 2020). Our constructions allow for *distribution-free selective inference* of black-box prediction intervals. This is a reasonable middle ground between fully marginal inference (standard conformal) and fully conditional inference (impossible). To avoid introducing an entirely new problem in the discussion, we include a few more details for the interested reader in Section D of the supplement.

References

Aharoni, E. and Rosset, S. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 76(4):771–794, 2014.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference (accepted)*, 2020.

Benjamini, Y. and Yekutieli, D. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

Benjamini, Y., Hochberg, Y., and Kling, Y. False discovery rate control in pairwise comparisons. *Research Paper, Department of Statistics and Operations Research, Tel Aviv University*, 1993.

Benjamini, Y., Hochberg, Y., and Stark, P. Confidence intervals with more power to determine the sign: Two ends constrain the means. *Journal of the American Statistical Association*, 93(441):309–317, 1998.

Foster, D. and Stine, R. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 70(2):429–444, 2008.

Gelman, A. and Tuerlinckx, F. Type s error rates for classical and bayesian single and multiple comparison procedures. *Computational Statistics*, 15(3):373–390, 2000.

Gelman, A., Hill, J., and Yajima, M. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012.

Goeman, J. J., Solari, A., and Stijnen, T. Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Statistics in medicine*, 29(20): 2117–2125, 2010.

Javanmard, A. and Montanari, A. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics*, 46(2):526–554, 2018.

Lei, J. and Wasserman, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76 (1):71–96, 2014.

Ramdas, A., Yang, F., Wainwright, M., and Jordan, M. Online control of the false discovery rate with decaying memory. In *Advances In Neural Information Processing Systems*, pp. 5655–5664, 2017.

Ramdas, A., Zrnic, T., Wainwright, M., and Jordan, M. SAFFRON: an adaptive algorithm for online control of the false discovery rate. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4286–4294, 2018.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

Stephens, M. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2016.

Storey, J. D. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

Tian, J. and Ramdas, A. ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls. In *Advances in Neural Information Processing Systems*, pp. 9388–9396, 2019.

Vovk, V. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490, 2012.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Weinstein, A. and Yekutieli, D. Selective sign-determining multiple confidence intervals with FCR control. *Statistica Sinica (to appear)*, 2014.

Weinstein, A., Fithian, W., and Benjamini, Y. Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association*, 108(501):165–176, 2013.