Supplement to "Online Control of the False Coverage Rate and False Sign Rate" by Weinstein & Ramdas.

A. Proofs

Proof of Lemma 1. Recall that we wish to prove

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

Without loss of generality, we can ignore the case when $S_i = 0$ almost surely for some i; in other words, if we would never select θ_i, then $V_i = 0$ almost surely, and we can just ignore the time instant i. Hence, we only consider the case when at least one value of X_i leads to selection.

To derive a bound on $\mathbb{E}[A_i]$, consider the following thought experiment. Let us hallucinate what selection decisions would have occurred under a slightly different series of observations, namely

$$\widetilde{X} := (X_1, X_2, \ldots, X_{i-1}, X^*, X_{i+1}, \ldots, X_T),$$

where X^* is any value that would have led to selection of θ_i, which is a predictable choice, because it can be made based on only the predictable selection rule $\S_i$. Let the sequence of selection decisions made by the same algorithm on $\widetilde{X}$ be denoted $\widetilde{S}_i$, the levels be denoted $\widetilde{\alpha}_i$, and the constructed intervals be $\widetilde{I}_i$. We then claim that

$$A_i \equiv \frac{S_i \mathbb{1}_{\theta_i \notin I_i}}{\sum_{j \le T} S_j} = \frac{S_i \mathbb{1}_{\theta_i \notin I_i}}{\sum_{j \le T} \widetilde{S}_j} =: \widetilde{A}_i,$$

where we have intentionally altered only the denominator. To see that the above equality holds, first note that if $S_i = 0$, then $A_i = \widetilde{A}_i = 0$. Then note that if $S_i = 1$, then $\widetilde{S}_i = S_i$ for all i. Indeed, because $X_j = \widetilde{X}_j$, for $j \le i-1$, the first $i-1$ selection decisions are identical by construction; then if $S_i = 1$ (and $\widetilde{S}_i = 1$ by construction), then $\mathcal{F}^i = \widetilde{\mathcal{F}}^i$, and so every future selection decision is also identical (and also the constructed CIs, at levels α_i). Hence,

$$\begin{aligned}
\mathbb{E}[A_i] = \mathbb{E}\left[\widetilde{A}_i\right] &\overset{(a)}{\le} \mathbb{E}\left[\frac{\mathbb{1}_{\theta_i \notin I_i}}{\sum_{j \le T} \widetilde{S}_j}\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\frac{1}{\sum_{j \le T} \widetilde{S}_j} \mathbb{E}\left[\mathbb{1}_{\theta_i \notin I_i} \,\middle|\, \widetilde{\mathcal{F}}^{n \setminus i}\right]\right] \\
&\overset{(c)}{\le} \mathbb{E}\left[\frac{\alpha_i}{\sum_{j \le T} \widetilde{S}_j}\right] \overset{(d)}{\le} \mathbb{E}\left[\frac{\alpha_i}{\sum_{j \le T} S_j}\right],
\end{aligned}$$

where inequality (a) holds because $S_i \le 1$, equality (b) follows because $\frac{1}{\sum_{j \le T} \widetilde{S}_j}$ is $\widetilde{\mathcal{F}}^{n \setminus i}$-measurable because $\widetilde{S}_i = 1$ by construction, inequality (c) holds by definition (2) of a marginal CI, and inequality (d) holds because $\widetilde{S}_j \ge S_j$ for all j by the monotonicity of selection rules. This completes the proof of the lemma. □

The above is a generalization of lemmas that have been proved in the context of online FDR control by (Javanmard & Montanari, 2018; Ramdas et al., 2017), 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 $\{R_i = 1\} \equiv \{P_i \le \alpha_i\}$ is obviously directly related to the false discovery event $\{P_i \le \alpha_i, i \in \mathcal{H}_0\}$. We will later see that online FCR control captures online FDR control as a special case.

Proof of Theorem 3. Note that the selection rule in Definition 2 can be rewritten as

$$\mathcal{S}_i(X_i, I_i) = 1 \iff X_i \in \{x : \mathcal{I}_i(x, \alpha_i) \subseteq K_{il} \text{ for some } l\}, \tag{9}$$

which defines a predictable selection rule because α_i are predictable. Thus, the procedure in Definition 2 is LORD-CI for a predictable selection rule. Because the CI rules $\mathcal{I}_i$ are monotone, and the α_i output by the LORD-CI algorithm are also monotone by construction, we conclude that the *selection rule* (9) is also monotone according to condition (4). Hence, the procedure in Definition 2 is now the LORD-CI procedure for a predictable *and monotone* selection rule, which controls the FCR by Theorem 2. The last step is to observe that a false localization event implies a false coverage event (but not necessarily the other way around), and hence $\text{FLR}(T) \le \text{FCR}(T) \le \alpha$. □

B. Further discussion of conditional CIs

In this section we provide a more formal treatment of the conditional approach. In this approach, a nominal $(1-\alpha)$ conditional CI is constructed for each selection. Note that it is often simpler to condition also on $\mathcal{F}^{i-1}$, that is, to design the conditional CI so that for all $a \in [0, 1]$,

$$\Pr\{\theta_i \notin \mathcal{I}_i(X_i, a) \mid \mathcal{F}^{i-1}, S_i = 1\} \le a. \tag{10}$$

We start with proving that constructing conditional CIs controls the mFCR at level α.

Theorem 4. *Constructing a $(1-\alpha)$ conditional CI after every selection ensures that $\forall T \in \mathbb{N}$, $\text{mFCR}(T) \le \alpha$.*

Proof. From the definition of a conditional CI it follows immediately that

$$\begin{aligned}
\mathbb{E}[V_i \mid S_i = 1] &= \mathbb{E}[I_{\theta_i \notin I_i} \mid S_i = 1] \\
&= \Pr\{\theta_i \notin I_i \mid S_i = 1\} \\
&\le \alpha.
\end{aligned}$$

Together with the fact that $\mathbb{E}[V_i \mid S_i = 0] = 0$, we have

$$\mathbb{E}[V_i \mid S_i] \leq \alpha \qquad \text{a.s.},$$

and hence,

$$\begin{aligned}\mathbb{E}\left[\sum_i V_i\right] &= \sum_i \mathbb{E}[V_i] = \sum_i \mathbb{E}[S_i V_i] \\ &= \sum_i \mathbb{E}[S_i \mathbb{E}[V_i \mid S_i]] \leq \sum_i \mathbb{E}[\alpha S_i] \\ &= \alpha \sum_i \mathbb{E}[S_i] = \alpha \mathbb{E}\left[\sum_i S_i\right].\end{aligned}$$

Rearranging the first and last displays above yields the desired result. □

Constructing conditional CIs at the nominal level ensures also that FCR is controlled. As a matter of fact, even the conditional expectation of FCP given that at least one selection is made,

$$\text{pFCR}(T) := \mathbb{E}\left[\text{FCP}(T) \,\middle|\, \sum_{i=1}^{T} S_i > 0\right],$$

is controlled when using conditional CIs. We call the above the *positive* FCR, in analogy to the positive FDR (Storey, 2003).

Theorem 5. *Constructing a* $(1-\alpha)$ *conditional CI after every selection ensures that*

$$\text{pFCR}(T) \leq \alpha \qquad \forall T \in \mathbb{N}.$$

Proof. Consider any sequence $(s_1, \ldots, s_T) \in \{0,1\}^T$ such that $\sum_i s_i > 0$. We have

$$\begin{aligned}&\mathbb{E}\left[\frac{\sum_i V_i}{\sum_i S_i} \,\middle|\, S_1 = s_1, \ldots, S_T = s_T\right] \\ &= \frac{1}{\sum_i s_i} \mathbb{E}\left[\sum_i V_i \,\middle|\, S_1 = s_1, \ldots, S_T = s_T\right] \\ &= \frac{1}{\sum_i s_i} \mathbb{E}\left[\sum_{\{i \leq T : s_i = 1\}} I_{\theta_i \notin I_i} \,\middle|\, S_1 = s_1, \ldots, S_T = s_T\right] \\ &= \frac{1}{\sum_i s_i} \sum_{\{i \leq T : s_i = 1\}} \Pr\{\theta_i \notin I_i \mid S_1 = s_1, \ldots, S_T = s_T\} \\ &\overset{(a)}{=} \frac{1}{\sum_i s_i} \cdot \\ &\quad \sum_{\{i \leq T : s_i = 1\}} \Pr\{\theta_i \notin I_i \mid S_1 = s_1, \ldots, S_{i-1} = s_{i-1}, S_i = 1\} \\ &\leq \frac{1}{\sum_i s_i} \sum_{\{i \leq T : s_i = 1\}} \alpha \\ &= \alpha,\end{aligned}$$

where equality (a) uses the fact that the selection decisions $(S_{i+1}, \ldots, S_T)$ are independent of X_i given $(S_1, \ldots, S_i)$ because the selection rules $\{\mathcal{S}_j\}$ are predictable. The original claim follows by taking expectation over the conditional distribution of $S_1, \ldots, S_T$ given that $\sum_{i=1}^{T} S_i > 0$. □

We immediately conclude that with conditional $(1-\alpha)$ CIs we also have

$$\text{FCR}(T) = \text{pFCR}(T) \cdot \Pr\left\{\sum_{i=1}^{T} S_i > 0\right\} \leq \text{pFCR}(T) \leq \alpha.$$

B.1. An inconsistency of conditional CIs

We mentioned two main points of criticism regarding the conditional approach. One was the potential difficulties in actually computing conditional CIs in general. The second point was incompatibility of conditional CIs when solving a localization problem, for example, when constructing follow-up CIs after multiple hypothesis testing. In this subsection we look more carefully into the latter.

In realistic situations where our (online) model might be applicable, it is almost always the case that the researcher has in mind a question of primary importance and one (or more) of secondary importance. In the motivating example from the Introduction, the management might be interested first in knowing the sign of the parameters θ_i (say positive or nonpositive), but would also like to supplement with confidence limits each parameter whose sign was classified. In general, it is common practice to answer the question of primary interest by running a multiple comparisons procedure, for example a multiple hypothesis testing rule or, as would apply to our example, a multiple sign-classification rule. Because the follow-up question is posed only if the first question was answered (e.g., we want a CI only if we were able to classify the sign), the conditional approach might appear as natural to use at the second stage. Nevertheless, the purpose of this section is to demonstrate that constructing conditional CIs after running a multiple comparisons procedure might lead to contradictions. Moreover, if one insists on conditional CIs, the price of "resolving" these incompatibilities might be a serious loss in power.

Before proceeding, we would like to explain why we view the issue of incompatibility as problematic. Consider for simplicity the offline setting and suppose that $X_i \sim N(\theta_i, 1)$. Suppose that a level-α FSR procedure yielded a subset of parameters θ_i whose signs are classified as positive or non-positive. For example, this subset may include θ_2 and θ_5, with the sign of θ_2 classified as positive, and the sign of θ_5 classified as nonpositive. Trivially, then, the (somewhat artificial) CIs given by $(0, \infty)$ for θ_2 and $(-\infty, 0]$ for θ_5—and similarly for all other θ_i whose sign was classified—control the FCR at α. In other words, there

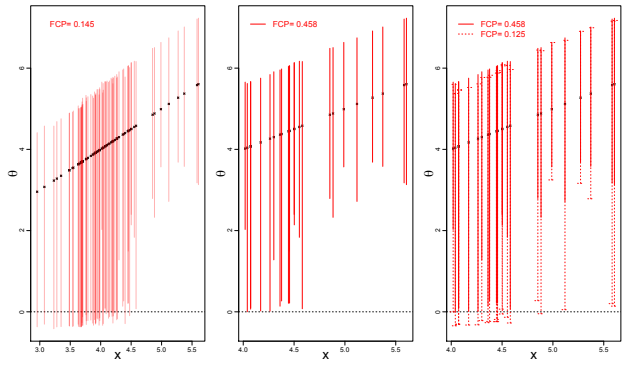

Figure 3. Conditional CIs following LORD++ selection. The left panel shows the 76 conditional 90% CIs originally constructed for rejected nulls. In the middle panel we kept only the 24 intervals that do not cover zero. The right panel shows these 24 intervals again (solid lines), but now along with their re-adjusted version (dashed lines). $2/3$ of the re-adjusted conditional CIs again cross zero.

exists at least one (trivial) CI protocol that controls the FCR and *is consistent with the FSR procedure*, in the sense that if the sign of θ_i was classified as positive (nonpositive) then the followup CI contains only positive (nonpositive) values. We therefore find it hard to justify a situation where one applies some valid FSR protocol, and follows up with CIs that provably control the FCR at the same level but 50% of them cross zero.

Let us now return to the motivating story of the Introduction, which we will here accompany with a simulation for illustration. Thus, we set $\alpha = 0.1$ and draw $m = 10,000$ parameters independently such that $\theta_i = (-1)^i \cdot 0.001$ (effectively "null") with probability 0.8, and $\theta_i = 2$ with probability 0.2. These represent the ground truth for the treatment effects of the first m drugs. The observations, which we assume arrive independently one at a time, are $X_i \sim N(\theta_i, 1)$. In the Introduction a CI was reported once X_i exceeded a fixed threshold. Suppose now that the statisticians are interested first in classifying the sign of a parameter as positive ("treatment effective") or nonpositive ("treatment ineffective"), and then follow up with CIs for those parameters whose sign was classified. To answer the first question, and being aware of multiplicity issues, the team decides to run the LORD++ testing procedure on two-sided p-values, where for each rejection they classify the sign as positive or nonpositive according as X_i is positive or nonpositive. This resulted in 76 selections in our simulation run, and makes sense as a criterion for whether to report an interval or not, because we know from the results of the current paper that the FSR is controlled. Furthermore, remember that, as shown in Section C, constructing the LORD-CI symmetric interval for each selected parameter, ensures at the same time control of the FCR and that none of the constructed CIs includes values of opposite signs. This is an output the management will, arguably, be content with seeing, at least in the sense that each reported CI is conclusive about the direction of the effect of the corresponding drug (because the intervals do not cross zero).

Instead, suppose that the statisticians will actually construct a 90% conditional CI for each selected parameter. Now, because we use conditional CIs, it is impossible to ensure that a constructed interval includes values of only one sign (that is, does not cross zero)—this is true no matter what choice we make for the conditional CI rule. Here we used the conditional CI of Weinstein et al. (2013) which inverts shortest acceptance regions. The left panel of Figure 3 shows the 76 constructed 90% conditional CIs. We know that this strategy controls the FCR (in our single realization of the experiment 14.5% of the constructed intervals are non-covering), but, less conveniently, there are also 52 (about 68%!) of these that cross zero. Hence, the team of statisticians will first have to reconcile the fact that on the one hand, each selected parameter can be safely classified for sign (as far as FSR is controlled), and on the other hand some intervals still include both positive and negative values. In any case—even if this incompatibility is overlooked—the management should certainly complain about the CIs that cross zero (because these are ambiguous about the direction of the effect of the corresponding drug). Trying to rectify the situation, they may ask to remove all CIs that do cross zero; unfortunately, doing this they would generally lose FCR control. The middle panel of Figure 3 shows the subset of (original) conditional CIs which do not cross zero; almost half of these (45.8%) do not cover their parameter. Nevertheless, the statisticians might propose at this stage to still keep only the CIs that do not cross zero, but re-adjust them for the fact that further selection took place, by constructing again conditional CIs with an appropriate cutoff. This will admittedly restore FCR control: the right panel of Figure 3 shows the re-adjusted CIs with dashed lines, and the proportion of such intervals that fail to cover their parameter drops again to 0.125. The problem is that some of the re-adjusted CI cross zero again (in fact, a much higher proportion than that in the first place), taking us back to the previous stage. If we were now to repeat the process by discarding the new 16 re-adjusted CIs that cross zero, we would be left with only 8 selections before even adjusting the CIs again. In other words, we are already down from the 76 sign-determining LORC-CI intervals to no more than 8 if we use conditional CIs. In general, this cycle could continue until there are very few parameters to report a CI for (maybe none). We should remark at this point that using a conditional CI that has better sign-determining properties, like the two options in Weinstein et al. (2013), could improve the results for the conditional approach, that is, we might

end up with more reported CIs. However, as we remarked before, the phenomenon in its essence remains regardless of the choice of the conditional CI.

C. An illustration of the Modified Quasi-Conventional (MQC) marginal CI

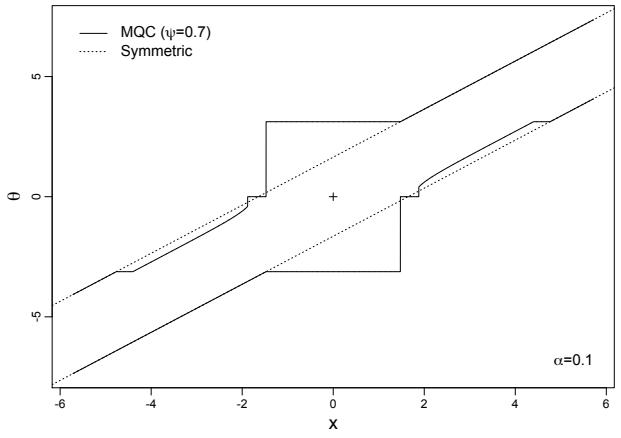

Figure 4. The Modified Quasi-Conventional (MQC) CI rule of Weinstein & Yekutieli (2014). Solid lines are lower and upper endpoints of the MQC CI for each observation value x. Dotted lines are lower and upper endpoints of the usual two-sided CI. It can be seen that the MQC interval excludes values of opposite signs earlier, that is at a smaller x value, than the usual two-sided CI. The parameter $\psi \in (0.5, 1)$ controls how early sign-determination occurs, and here $\psi = 0.7$ is used. The unusual constant shape of the MQC in the neighborhood of zero does not matter because intervals that cross zero are anyway discarded in a sign-determining selective-CI procedure.

D. Selective conformal inference

Conformal prediction is a general nonparametric technique for producing marginally valid prediction intervals under almost no regularity assumptions on the data generating process beyond exchangeability of the data. A simple version of the setup can be explained as follows. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be drawn i.i.d. from some joint distribution $P_{XY} = P_X \times P_{Y|X}$, which are supported on the domain $\mathcal{X} \times \mathcal{Y}$, where for simplicity let $\mathcal{Y} = \mathbb{R}$. Given a test point X_{n+1} drawn i.i.d. from P_X, our task is to provide a prediction interval for the unobserved Y_{n+1}.

Conformal prediction begins by hallucinating a value y, to form a new dataset $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$. One may then train any regression algorithm $f : \mathcal{X} \to \mathcal{Y}$ on this set of $n + 1$ points to obtain $\widehat{f}$, and calculate the $n + 1$ in-sample residuals $r_i = Y_i - \widehat{f}(X_i)$ for $i \in [n]$ and $r_{n+1} = y - \widehat{f}(X_{n+1})$. We then "reject" y if r_{n+1} is in the largest α-quantile of all $n + 1$ residuals. We then repeat this whole process for every possible $y \in \mathcal{Y}$. The final prediction interval $\mathcal{I}(X_{n+1}, \alpha)$ consists of all those ys that we did not reject. The intuition is that when $y = Y_{n+1}$, then all $n+1$ residuals are exchangeable, and so the rank of r_{n+1} among $r_1, \ldots, r_{n+1}$ is uniform. Hence the probability of rejecting $y = Y_{n+1}$ and excluding it from the interval, equals the probability that r_{n+1} is in the largest α-quantile of $r_1, \ldots, r_{n+1}$, which is at most α. The formal guarantee is that $I_{n+1} := \mathcal{I}(X_{n+1}, \alpha)$ is marginally valid:

$$\Pr\{Y_{n+1} \notin \mathcal{I}(X_{n+1}, \alpha)\} \leq \alpha,$$

where the probability is taken over all $(n + 1)$ draws from P_{XY}. Remarkably, this guarantee holds with no assumptions on the distribution P_{XY} or on the regression algorithm f (these may affect the length of the intervals, but not their validity). However, a conditional conformal guarantee is in general impossible, meaning that if we do not make any distributional assumptions and we would like a guarantee of the form

$$\Pr\{Y_{n+1} \notin \mathcal{I}(X_{n+1}, \alpha) | X_{n+1} = x\} \leq \alpha$$

to hold for any x, then the corresponding conditional conformal interval $\mathcal{I}(X_{n+1}, \alpha)$ must have infinite length. The impossibility of fully conditional conformal prediction was pointed out by Vovk (2012), elaborated further by Lei & Wasserman (2014) and Barber et al. (2020).

The relationship of the above discussion to the current paper is as follows. There was nothing in particular that restricted the setup of the current paper to confidence intervals for parameters θ_i based on observations X_i. The setup just as easily covers prediction intervals for outcomes Y_i based on features X_i. To understand the implications, suppose we were to observe a sequence $X_{n+1}, \ldots, X_{n+m}, \ldots$ of test points drawn i.i.d. from P_X, and we do not wish to cover all of the corresponding Ys but just some subset of them. Then, one may construct marginally valid prediction intervals (at predictable levels α_i using LORD-CI) for an adaptively selected subset of X_is, and this paper provides an FCR control guarantee on those selected intervals.

Hence, even though conditional conformal inference is impossible, our works implies that "selective conformal inference" is possible. There is no contradiction here: an FCR guarantee is weaker than a conditional guarantee. Also, the FCR guarantee cannot really be used to get a conditional guarantee; indeed, if one was to only select X_i for coverage if it is in a very small ϵ-ball around a given point x (to approximate the conditional coverage guarantee), then such selections would be very infrequent, and the α_i used would be very close to zero, resulting in an exceedingly wide interval. As $\epsilon \to 0$, we would also see $\alpha_i \to 0$, and thus the length of the selected interval would become infinite.

We end this section by remarking that nothing was particular to conformal prediction intervals; the FCR guarantees would also apply to any other marginally-valid prediction intervals. Further, there was also nothing particular to the online setting; indeed, such FCR control can also be guaranteed in the offline setting.