Online multiple testing with e-values

Ziyu Xu* Aaditya Ramdas†

November 7, 2023

Abstract

A scientist tests a continuous stream of hypotheses over time in the course of her investigation — she does not test a predetermined, fixed number of hypotheses. The scientist wishes to make as many discoveries as possible while ensuring the number of false discoveries is controlled — a well recognized way for accomplishing this is to control the false discovery rate (FDR). Prior methods for FDR control in the online setting have focused on formulating algorithms when specific dependency structures are assumed to exist between the test statistics of each hypothesis. However, in practice, these dependencies often cannot be known beforehand or tested after the fact. Our algorithm, e-LOND, provides FDR control under arbitrary, possibly unknown, dependence. We show that our method is more powerful than existing approaches to this problem through simulations. We also formulate extensions of this algorithm to utilize randomization for increased power, and for constructing confidence intervals in online selective inference.

Contents

*Department of Statistics and Data Science, Carnegie Mellon University, USA. Email: `xzy@cmu.edu`.

†Departments of Statistics and Data Science, and Machine Learning, Carnegie Mellon University, USA. Email: `aramdas@cmu.edu`.

1 Introduction

Science advances one hypothesis at a time. Moreover, the rate at which new hypotheses are tested has drastically increased in recent decades to the point where a single scientist can quickly test hundreds to thousands of hypotheses with the aid of computation. For example, a geneticist can now sequence thousands of genes from trial subjects and individually determine whether each of these genes has an effect on phenotypes of interest (e.g., disease, physical characteristics, etc.). A team of data scientists can test many variations of a website or app in A/B experiments to determine which version maximizes desirable user metrics. The key feature of all these examples is that hypotheses are being formulated and tested in an *online* fashion — the total number of hypotheses that are tested is unknown beforehand and possibly infinite. Thus, we can formulate the online multiple testing problem, as receiving a stream of hypotheses, $H_1, H_2, \ldots$ — typically, these are the null hypotheses we wish to reject (e.g., this gene has no effect on this disease, there is no association between socioeconomic status and future earning potential, this recommendation algorithm does not increase average user view count, etc.). A subset, $\mathcal{H}_0 \subseteq \mathbb{N}$, of these null hypotheses are truly null, where $\mathbb{N}$ denotes the natural numbers. We wish to discover all the hypotheses that are not null, i.e., discover the non-null hypotheses $\mathcal{H}_1 \coloneqq \mathbb{N} \setminus \mathcal{H}_0$. For each hypothesis, we observe some data and must immediately decide whether it is a discovery or not before observing future hypotheses. Thus, we denote the set of discoveries so far as $\mathcal{R}_1 \subseteq \mathcal{R}_2 \subseteq \cdots \subseteq \mathbb{N}$. The *false discovery proportion* (FDP) refers to the proportion of discoveries in a discovery set $\mathcal{R}$ that are truly null. We want to control the *false discovery rate* (FDR), which is the expectation of the FDP. Define these as follows.

$$\mathrm{FDP}(\mathcal{R}) \coloneqq \frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}, \qquad \mathrm{FDR}(\mathcal{R}) \coloneqq \mathbb{E}[\mathrm{FDP}(\mathcal{R})].$$

$(X_t)_{t \in \mathbb{I}}$ denotes a sequence of objects indexed by a set $\mathbb{I}$ — we drop the index set and write (X_t) it is clear from context (often $\mathbb{N}$). Our goal is to produce discovery sets $(\mathcal{R}_t)$ that satisfy the following guarantee:

$$\mathrm{FDR}(\mathcal{R}_t) \leq \alpha \text{ for all } t \in \mathbb{N}, \tag{1}$$

while maximizing the number of discoveries. FDR is reasonable metric to control in applications where one wishes to filter candidates that are promising before doing more extensive follow-up studies, e.g., clinical trials for drugs, genome-wide association studies for genetic factors, features for pushing to production, etc. We elaborate on the motivations for considering the FDR error metric in Appendix A.1. Robertson et al. (2023) comprehensively surveys the existing literature of online multiple testing. In particular, multiple previous works have devoted significant effort to formulating different types of dependency that can arise in natural situations and deriving algorithms that provide online FDR control under these dependence structures (Zrnic et al., 2021, 2020; Fisher, 2022b,a). These works have considered dependencies that are natural to the online setting (i.e., local dependence and dependence between asynchronously initiated experiments) as well as the popular PRDS condition (Benjamini and Yekutieli, 2001). However, under *unknown or arbitrary* dependence in the data, the assumptions for these algorithms are violated and they do not provably control the FDR.

There are many circumstances where one wishes to be robust to arbitrary dependence — we list some below:

- **Data reuse**. A natural way in which unknown dependency might arise is when one uses the same dataset to evaluate a large number of hypotheses. Although reusing data for different hypothesis tests is not generally a statistically valid practice, this practice inevitably occurs, as data collection may be difficult or prohibitively expensive. For example, in many applied areas of machine learning, the same dataset may be used to evaluate many different methods, e.g., Kaggle competitions (Bojer and Meldgaard, 2021), the UCI data repository (Newman et al., 1998). Similarly, open data repositories in science also are reused across many studies (1000 Genomes Project Consortium, 2015; Wellcome Trust Case Control Consortium, 2007; Koscielny et al., 2014). Data reuse naturally comes up in offline policy evaluation in reinforcement learning, since often deploying a new policy has costs (e.g., expenses incurred by new actions, loss of revenue if a policy underperforms, etc.), and one would wish to backtest many policies on previously collected data. In all these cases, the statistics calculated for each test are highly dependent, since they use the same data.

- **Temporal overlap**. This type of dependency is considered primarily in works involving local dependencies (Zrnic et al., 2021), as it occurs when data collected for different hypotheses overlap or are subject to temporal noise. For example, in A/B testing, users are incrementally added to each experiment over time. However, since there is no partitioning of users across experiments, experiments may overlap in users. This induces a dependence among the resulting test statistics. Temporal events (e.g., holidays or weekends) can also induce time-dependent noise. We elaborate on the "doubly sequential framework" relevant to this setting in Section 2.

- **Inherent dependence**. Dependence between statistics might simply arise because of the data generating process. One common type is dependence that arises from sampling without replacement (WoR) from a finite population. Sampling WoR naturally arises when we wish to test the average treatment effect of a treatment on the finite population (Splawa-Neyman et al., 1990) — the statistics calculated for different treatments allocated to different samples are dependent — we simulate our methods in this setting in Section 5. Similarly, dependence also arises when doing coarser cluster (rather than individual) based randomization (Campbell et al., 2007). Dependence can also come from a data-dependent sampling mechanism, which we can observe in multi-armed bandits or adaptive sampling settings.

In many experiments, one may not know ahead of time which combination of the aforementioned types of dependencies may occur, nor the specific structure they may take. This is particularly relevant in online multiple testing, since the nature of the hypotheses being tested and which types of data are being used to conduct the tests are not known a priori. Hence, being simultaneously powerful and robust to arbitrary dependence is a highly practical desiderata.

The primary of contribution of this paper is a new algorithm, e-LOND, that provably controls FDR, i.e., satisfies (1), under unknown and arbitrary dependence, while being more powerful (i.e, makes more discoveries) than previous state-of-the-art algorithms. Our method accomplishes this by utilizing *e-values*, a class of statistics that has garnered significant recent attention in hypothesis testing. E-values are central in sequential testing (Ramdas et al., 2020, 2021) as every admissible sequential test utilizes an e-value. We characterize a "doubly sequential framework" of scientific experimentation that combines sequential tests with online multiple testing in Section 2, and illustrate how retaining validity under arbitrary dependence is particularly useful in this framework. A notable example of an e-value is the universal inference statistic (Wasserman et al., 2020), which allows for testing of composite nulls without regularity conditions. This, in turn, enables the construction of tests for novel problems where no prior valid test exists — an example of this is testing whether a distribution is log-concave (Dunn et al., 2022; Gangrade et al., 2023). The kinds of hypotheses for which e-values are applicable is quite comprehensive. We refer the reader to Ramdas et al. (2023) for thorough collection of examples for which e-values are applicable.

P-values vs. e-values. Since the formulation of online multiple testing by Foster and Stine (2008), solutions have only assumed a *p-value*, P_t, is associated with hypothesis H_t and satisfies the following,

$$\mathbb{P}\left(P_t \leq s\right) \leq s \text{ for all } s \in [0,1] \text{ if } t \in \mathcal{H}_, \tag{2}$$

for all $t \in \mathbb{N}$. We consider the novel setting where, instead, an *e-value*, E_t, accompanies each hypothesis H_t and satisfies the following property for all $t \in \mathbb{N}$:

$$\mathbb{E}[E_t] \leq 1 \text{ if } t \in \mathcal{H}_0. \tag{3}$$

An *online multiple testing algorithm* is a sequence of (possibly random) test levels (α_t), where $\alpha_t \in [0,1]$ for all $t \in \mathbb{N}$, and the algorithm produces discovery set $\mathcal{R}_t$ at the tth step in the following fashion:

$$\mathcal{R}_t = \begin{cases} \{i \in [t] : P_i \leq \alpha_i\} \text{ if using p-values,} \\ \{i \in [t] : E_i \geq 1/\alpha_i\} \text{ if using e-values} \end{cases}.$$

The definition of $\mathcal{R}_t$ in the e-value case is equivalent to the p-value case if we assumed our p-values were formulated as $P_t = 1/E_t$ — one can see this is a bona fide p-value by applying Markov's inequality to the e-value definition in (3). One can consider e-value algorithms as operating on a special type of p-values. We leverage the specific properties of e-values to derive more powerful algorithms that remain valid even under arbitrary dependence.

Our contributions. We make the three following contributions in the main paper.

1. *Powerful online FDR control under arbitrary dependence with e-values.* The current method for online FDR control under arbitrary dependence, the r-LOND algorithm (Javanmard and Montanari, 2018; Zrnic et al., 2021), is unnecessarily conservative when applied to e-values. The r-LOND algorithm corrects each of its test levels by an additional factor that is logarithmic in the number of hypotheses tested so far, compared to its counterpart, the LOND algorithm, that ensures FDR control under a much more stringent assumption of positive dependence. This is similar to the penalty paid by the Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001) in the offline setting. Our algorithm, e-LOND, operates on e-values, but does not require the additional correction. Thus, it can maintain FDR control regardless of the dependence structure and dominates the standard r-LOND algorithm. Another previous approach to FDR control under dependence is the LORD* algorithm, which requires a priori knowledge of which hypotheses have statistics that are dependent. Our numerical simulations in Section 5 show that e-LOND is more powerful than r-LOND and becomes more powerful than LORD* when more hypotheses are mutually dependent.

2. *Additional power through randomization.* If one is interested in maximizing the power of their online multiple testing procedure, then one can incorporate randomization in the manner of of Xu and Ramdas (2023), who use randomization to improve offline multiple testing procedures. We develop variants of e-LOND and r-LOND (Ue-LOND and Ur-LOND, respectively), that use the randomization of a single uniform random variable to increase their power over their deterministic counterparts. These randomized methods dominate (i.e., never make fewer, and often make more discoveries) their deterministic versions and hence should be employed if one is interested in making as many discoveries as possible.

3. *Online FCR control with no restrictions on selection rules or dependence on e-CIs.* In addition to online FDR control, we also provide novel results for the online selective confidence interval (CI) problem introduced by Weinstein and Ramdas (2020). In this problem, one wishes to output, in an online fashion, CIs for a stream of parameters such that the overall false coverage rate (FCR) of all the CIs is controlled. This problem adds in the additional complexity of having a selection rule — while a discovery is made at the tth hypothesis solely based on its test level α_t, one decides whether a parameter should be selected for CI construction based on a selection rule $\mathbf{S}_t$ (which uses the observed data for the current and past parameters) that is separate from the coverage level of the CI, $1 - \alpha_t$. The extension of e-LOND to the online selective CI problem can control FCR under any sequence of selection rules, and arbitrary dependence. The sole caveat of this algorithm is that it operates on a subset of CIs based on e-values, called e-CIs (Vovk and Wang, 2023; Xu et al., 2022), which have been used for offline FCR control.

Outline. In Section 2, we discuss the "doubly sequential" framework that abstracts scientific experimentation. We recap existing online multiple testing algorithms and introduce the e-LOND algorithm in Section 3. In Section 4 we devise methods for the online selective inference problem from e-LOND. We demonstrate the power of e-LOND empirically through numerical simulations in Section 5, and summarize our findings in Section 9. We defer discussion of related work to Section 7 and appendix A.2. Further, we apply our methods to an online version of the model-free selective inference problem of Jin and Candès (2023) in Section 6, and evaluate their performance on real data from protein prediction task. Lastly, we show a sharpness result on the FDR control of e-LOND in Appendix C, i.e., there exists instances where the true FDR is arbitrarily close to α.

2 Doubly sequential inference

E-values are particularly applicable to the sequential fashion in which data is gathered in many modern applications of hypothesis testing. In the sequential setting, samples are received one at a time, e.g., patients entering a clinical trial, users joining an A/B test, etc. To maximize efficiency, we collect data samples $X_1, X_2, \ldots$ (here we are indexing by sample, rather than hypothesis) and stop sampling as soon as we are able to make a decision about the result of the experiment. A key concept for sequential testing is the *e-process*, which is a process (M_t) where M_t is a function of the first t samples $(X_1, \ldots, X_t)$ and satisfies the following property:

$$\mathbb{E}[M_\tau] \leq 1 \text{ for all stopping times } \tau \text{ under } H_0. \tag{4}$$

A *stopping time* is a random time τ that can be determined based on the data seen so far, i.e., one can determine whether $\tau = t$ solely by using $(X_1, \ldots, X_t)$. From the definition of an e-process in (4), one can see M_τ is an e-value, so making a discovery when $M_\tau \geq \alpha^{-1}$ is a valid hypothesis test with Type I error of at most α. Consequently, a ubiquitous stopping time is the first time at which M_t exceeds the test threshold α^{-1}. Ramdas et al. (2020) showed that any admissible sequential test which allows early stopping of this sort *must* be derived from e-processes, making e-values a central and necessary component of sequential testing.

This leads us to the *doubly sequential framework* (Robertson et al., 2023), where both samples are hypotheses arrive sequentially, as a widely applicable framework for how scientific experimentation is

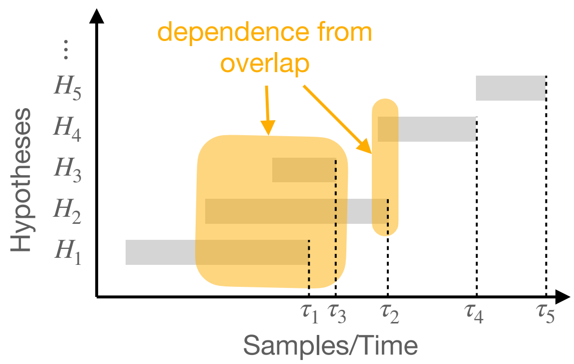

Figure 1: A cartoon of the doubly-sequential framework for experimentation. Real world time or the number of samples collected is on the x-axis — experiments run sequentially, stopping at τ_t when enough samples are collected for hypothesis H_t. Hypotheses arriving in a stream are shown on the y-axis. As a result of the overlap in time that data for different experiments is collected, dependence between hypotheses can occur.

done. An online multiple testing algorithm that utilizes e-values a quite useful in this framework, since e-values are critical to sequential testing. Figure 1 illustrates this concept. Both data and hypotheses arrive in streams, and one must be able to test new hypotheses and utilize new data as evidence in an online fashion. When many experiments are run simultaneously, the data gathered for each experiment are *dependent*, either due noise that jointly affects samples collected at a similar time (e.g., season fluctuations affecting the e-commerce habits of users), or because some experiments might share some of the collected data (e.g., clinical endpoints that utilize data from prior trials). Thus, a common application of this framework is in large scale A/B testing at companies (Xu et al., 2015), where separate data scientists are starting new experiments regularly, and have concurrent existing experiments that gather data sequentially. Yang et al. (2017) illustrate an instance where the data for each hypothesis is collected through a multi-armed bandit.

Regardless, all these scenarios can involve complicated and unknown dependence between the statistics for testing each hypothesis. Thus, our methods that are robust to dependence allow for valid inference in the doubly sequential framework, and we verify this empirically in our experiments in Section 5.

3 e-LOND: FDR control via e-values

To prepare ourselves for e-LOND, we first recap what the current state-of-the-art algorithms are. Let a *discount sequence* (γ_t) be a fixed sequence of nonnegative reals that satisfy $\sum_{t=1}^{\infty} \gamma_t \leq 1$, and $\alpha \in [0, 1]$ is our desired level of FDR control. For all sequences of discovery sets $(\mathcal{R}_t)$, we let $\mathcal{R}_0 = \emptyset$. An algorithm that produces a sequence of discovery sets $(\mathcal{R}_t^1)$ *strictly dominates* an algorithm that produces $(\mathcal{R}_t^2)$ iff (1) $\mathcal{R}_t^1 \supseteq \mathcal{R}_t^2$ on all sequences of p-values (P_t) (or e-values (E_t)) and all $t \in \mathbb{N}$, and (2) there is a sequence p-values (P_t) (or e-values (E_t)) s.t. there exists $t \in \mathbb{N}$ where $\mathcal{R}_t^1 \supset \mathcal{R}_t^2$. We define this as *strictly dominates in expectation* if condition (2) is instead the existence of $t \in \mathbb{N}$ such that $\mathbb{E}[|\mathcal{R}_t^1| \mid (E_t)] > \mathbb{E}[|\mathcal{R}_t^2| \mid (E_t)]$, i.e., the expected number of discoveries is strictly larger when taken only over the randomness in the algorithm.

3.1 Prior work: the LOND and r-LOND algorithms

We first recall the LOND algorithm. For each $t \in \mathbb{N}$ define:

$$\alpha_t^{\mathrm{LOND}} \coloneqq \alpha\gamma_t \cdot (|\mathcal{R}_{t-1}^{\mathrm{LOND}}| + 1),$$

where $(\mathcal{R}_t^{\mathrm{LOND}})$ are the corresponding discovery sets. The LOND algorithm requires p-values to be independent or positively dependent for FDR control.

Fact 1 (Theorem 4 (Zrnic et al., 2021))**.** *For p-values* (P_t) *that satisfy* (2) *and are independent or PRDS (Zrnic et al., 2021, Definition 1),* $\mathrm{FDR}(\mathcal{R}_t^{\mathrm{LOND}}) \leq \alpha$ *for each* $t \in \mathbb{N}$*.*

To achieve FDR control under arbitrary dependence, the r-LOND algorithm outputs more conservative test levels. For each $t \in \mathbb{N}$, define

$$\alpha_t^{\text{r-LOND}} \coloneqq \alpha\gamma_t \cdot \beta_t(|\mathcal{R}_{t-1}^{\text{r-LOND}}| + 1). \tag{5}$$

Here, (β_t) is a sequence of *reshaping functions* (Blanchard and Roquain, 2008). A reshaping function $\beta : [0, \infty) \mapsto [0, \infty)$ is a nondecreasing function that can be written in the form $\beta(r) = \int_0^r x d\nu(x)$ where ν is any probability measure on $[0, \infty)$. Let $(\mathcal{R}_t^{\text{r-LOND}})$ denote the sequence of discovery sets output by r-LOND.

Fact 2 (Theorem 2.7 (Javanmard and Montanari, 2015), Theorem 4 (Zrnic et al., 2021) [1])**.** *Under arbitrary dependence in* (P_t)*, i.e., under* (2)*,* $\mathrm{FDR}(\mathcal{R}_t^{\text{r-LOND}}) \leq \alpha$ *for each* $t \in \mathbb{N}$*.*

A typical choice of reshaping function is

$$\beta_t^{\mathrm{BY}}(r) = (\lfloor r \rfloor \wedge t)/\ell_t,$$

where $\ell_t \coloneqq \sum_{i=1}^t 1/i$ — this is the choice used by the Benjamini-Yekutieli (BY) procedure (Benjamini and Yekutieli, 2001) for offline FDR control. Hence, one can consider LOND and r-LOND as the online analogs of the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) for independent or PRDS p-values and the BY procedure for arbitrarily dependent p-values, respectively.

3.2 The e-LOND algorithm

Our e-LOND *algorithm* achieves the best-of-both worlds in the sense it has the same powerful test levels as LOND, but also is valid under arbitrary dependence like r-LOND. For each $t \in \mathbb{N}$, define

$$\alpha_t^{\text{e-LOND}} \coloneqq \alpha\gamma_t \cdot (|\mathcal{R}_{t-1}^{\text{e-LOND}}| + 1).$$

$(\mathcal{R}_t^{\text{e-LOND}})$ denotes the resulting discovery sets. The following is our main result.

Theorem 1. *Under arbitrary dependence on e-values* (3)*,* $\mathrm{FDR}(\mathcal{R}_t^{\text{e-LOND}}) \leq \alpha$ *for each* $t \in \mathbb{N}$*. In addition,* e-LOND *strictly dominates* r-LOND *applied to* $(1/E_t)$ *for any sequence of reshaping functions* (β_t)*.*

The proof relies on a simple observation about any e-value E and test level $\alpha \in [0, 1]$ that allows us to directly upper bound the indicator of whether a discovery is made or not by the e-value itself:

$$\mathbf{1}\left\{E \geq \alpha^{-1}\right\} = \mathbf{1}\left\{\alpha E \geq 1\right\} \leq \alpha E. \tag{6}$$

---

[1] Strictly speaking, r-LOND in Zrnic et al. (2021) is formulated as $\alpha_t^{\text{r-LOND}} = \alpha\gamma_t \cdot \beta_t(|\mathcal{R}_{t-1}| \vee 1)$ which is less powerful than (5), the latter being the original r-LOND (Javanmard and Montanari, 2015). However, the proofs of Zrnic et al. (2021) carry through to the original r-LOND.

We defer the full proof to Section 8.1. Further, we show in Appendix C that this level of FDR control is *sharp*, i.e., one can design instances of e-values where the true FDR of e-LOND is arbitrarily close to the upper bound of α.

The e-LOND algorithm has the same test levels (α_t) as LOND, but we use different notation to emphasize that e-LOND operates on e-values with no restrictions on dependence and LOND operate on p-values that are independent or satisfy PRDS. This is similar to the relationship between the e-BH procedure (Wang and Ramdas, 2022) and BH for offline FDR control.

In addition, we can show r-LOND is actually a special case of e-LOND. To clarify how r-LOND is subsumed by e-LOND under arbitrary dependence, we introduce the notion of calibration. Any p-value P can be calibrated into an e-value $E = f(P)$ using a *calibrator* (Vovk and Wang, 2021). A calibrator $f : [0, 1] \mapsto [0, \infty)$ is an nonincreasing, upper semicontinuous function that satisfies $\int_0^1 f(x)dx \leq 1$. We can define a specific sequence of calibrators (f_t) that transform p-values into e-values such that r-LOND is a special case of e-LOND.

Corollary 1. *If p-values (P_t) satisfy (2), we can construct an e-value $E_t = f_t(P_t)$ for each $t \in \mathbb{N}$ from a sequence of calibrators (f_t). We achieve $\mathrm{FDR}(\mathcal{R}_t^{\text{e-LOND}}) \leq \alpha$ for each $t \in \mathbb{N}$ by Theorem 1. If we define f_t as follows:*

$$f_t(p) = (\alpha\gamma_t \cdot \lceil (p\ell_t/(\alpha\gamma_t)) \vee 1 \rceil)^{-1}$$

we recover r-LOND *for FDR control under arbitrary dependence described in Fact 2. This allows us to reap the benefits of* e-LOND *when only some hypotheses may have e-values, and the rest have p-values — we can calibrate just the p-values before running* e-LOND.

3.3 More power through randomization

Building on recent advances by Xu and Ramdas (2023) for offline multiple testing, we can strictly improve both e-LOND and r-LOND by incorporating independent randomization. Let E be an e-value and $\widehat{\alpha} \in [0, 1]$ be a possibly random threshold that may depend on E. Let U be a uniform random variable on $[0, 1]$ that is independent of both E and $\widehat{\alpha}$. Define the following randomized e-value:

$$S_{\widehat{\alpha}}(E) \coloneqq (E \cdot \mathbf{1}\left\{E \geq \widehat{\alpha}^{-1}\right\}) \vee (\mathbf{1}\left\{U \leq E\widehat{\alpha}\right\} \widehat{\alpha}^{-1}),$$

Fact 3 (Proposition 2 of Xu and Ramdas 2023). *$S_{\widehat{\alpha}}(E)$ is also an e-value. Further, note that*

$$\mathbf{1}\left\{S_{\widehat{\alpha}}(E) \geq \widehat{\alpha}^{-1}\right\} = \mathbf{1}\left\{E \geq \widehat{\alpha}^{-1} \cdot U\right\}$$

We now define Ue-LOND, a randomized version of e-LOND. Let (U_t) be a sequence of uniform random variables on $[0, 1]$ that are independent of (E_t).

$$\alpha_t^{\text{Ue-LOND}} \coloneqq \alpha_t^{\text{e-LOND}} \cdot U_t^{-1}. \tag{7}$$

Let $(\mathcal{R}_t^{\text{Ue-LOND}})$ be the sequence of discovery sets output by Ue-LOND. The following is our second main result.

Theorem 2. *Under arbitrary dependence on e-values (3), $\mathrm{FDR}(\mathcal{R}_t^{\text{Ue-LOND}}) \leq \alpha$ for each $t \in \mathbb{N}$. Further,* Ue-LOND *strictly dominates* e-LOND*in expectation.*

Proof. Ue-LOND in (7) is equivalent to applying Ue-LOND to $(S_{\alpha_t^{\text{e-LOND}}}(E_t))$. Hence, FDR control holds by Theorem 1. The domination is because $U_t^{-1} > 1 + \varepsilon$ with nonzero probability for all $\varepsilon > 0$, and is independent from (E_t). □

Note that (U_t) can all be equal, i.e., $U_1 = \cdots = U_t$, or they can be drawn independently for each hypothesis. To improve r-LOND, we use the following result.

Fact 4 (Lemma 1 Xu and Ramdas (2023))**.** *Let P be a superuniform random variable that can be arbitrarily dependent on a positive random variable R. Let U be a superuniform random variable that is independent of both P and R. Let c be a nonnegative constant and β be a reshaping function. Then, the following holds:*

$$\mathbb{E}\left[\frac{\mathbf{1}\left\{P \leq c\beta(R/U)\right\}}{R}\right] \leq c.$$

We can define the Ur-LOND procedure as follows:

$$\alpha_t^{\text{Ur-LOND}} \coloneqq \alpha\gamma_t\beta_t((|\mathcal{R}_{t-1}| + 1)/U_t),$$

with $(\mathcal{R}_t^{\text{Ur-LOND}})$ being the resulting discovery sets. We now present our third main result.

Theorem 3. *Under arbitrary dependence on p-values* (2)*,* $\text{FDR}(\mathcal{R}_t^{\text{Ur-LOND}}) \leq \alpha$ *for each* $t \in \mathbb{N}$*. Further,* Ur-LOND *strictly dominates* r-LOND *in expectation for reshaping functions* (β_t^{BY})*.*

We defer the proof to Section 8.2.

Corollary 2. *If we use reshaping function* β_t^{BY}*,* Ur-LOND *produces the following test levels:*

$$\alpha_t^{\text{Ur-LOND}} = \alpha\gamma_t(\lfloor(|\mathcal{R}_{t-1}^{\text{Ur-LOND}}| + 1)/U\rfloor \wedge t)/\ell_t.$$

Thus, by utilizing randomization, we are able to derive FDR controlling procedures that are never worse than their deterministic counterparts.

4 Online FCR control with e-CIs

Often, a scientist wishes not only to test the significance of an effect but also to measure the strength of the effect. Instead of receiving hypotheses in a stream, a scientist can consider a stream of parameters $\theta_1 \in \Theta_1, \theta_2 \in \Theta_2, \ldots$, but wishes to estimate only some of them, e.g., only ones that show signficiant positive effect. Here, we desire our selected CIs to be accurate in aggregate, i.e., we want to control the false coverage rate (FCR) — this problem was introduced by Weinstein and Ramdas (2020) as the *the online selective-CI problem*. For the tth parameter, the scientist receives some data (e.g., the results of an experiment) $X_t \in \mathcal{X}_t$ and designs a selection rule $\mathbf{S}_t : \mathcal{X}_t \mapsto \{0, 1\}$ to decide whether CI should be constructed for θ_t. If a parameter is selected, one must choose an error level $\alpha_t \in (0, 1)$ and construct a $(1 - \alpha_t)$-CI for θ_t. Let $S_t = \mathbf{S}_t(X_t)$ be an indicator variable that is 1 iff θ_t is selected for CI construction. We assume that one has access to a CI constructor $C_t : \mathcal{X}_t \times [0, 1] \mapsto 2^{\Theta_t}$ for each $t \in \mathbb{N}$ where $C_t(X, \alpha)$ satisfies the following property:

$$\mathbb{P}\left(\theta_t \notin C_t(X_t, \alpha)\right) \leq \alpha \text{ for every } \alpha \in [0, 1]. \tag{8}$$

Formally, the *false coverage proportion* (FCP), and the *false coverage rate* (FCR) are defined as follows:

$$\text{FCP}(\mathcal{S}_t) \coloneqq \sum_{i \in \mathcal{S}_t} \frac{\mathbf{1}\left\{\theta_i \notin C_i(\alpha_i)\right\}}{|\mathcal{S}_t| \vee 1},$$
$$\text{FCR}(\mathcal{S}_t) \coloneqq \mathbb{E}\left[\text{FCP}(\mathcal{S}_t)\right].$$

The methods of Weinstein and Ramdas (2020) relied on two key assumptions. The first is an explicit assumption on the dependence between hypotheses, i.e., X_t were independent or that $C_t(X_t, \alpha_t)$ is still a valid $(1 - \alpha_t)$-CI conditional on past selection decisions. The second is a restrictive monotonicity assumption on the selection rules $\mathbf{S}_t$. In Algorithm 1, we devise versions of e-LOND and Ue-LOND for

Algorithm 1: The e-LOND-CI and Ue-LOND-CI algorithms ensure FCR $\leq \alpha$ with no restrictions on the dependence between data (X_t) or the selection rules $(\mathbf{S}_t)$. Let (U_t) uniform random variables on $[0, 1]$ and independent of (X_t).

Input: E-CI constructors (C_t), discount sequence (γ_t), and FCR control level α.
for each $t \in \mathbb{N}$ **do**
  if running e-LOND-CI **then**
    $\alpha_t \coloneqq \alpha\gamma_t(|\mathcal{S}_{t-1}| + 1)$.
  else if running Ue-LOND-CI **then**
    $\alpha_t \coloneqq \alpha\gamma_t(|\mathcal{S}_{t-1}| + 1) \cdot U_t^{-1}$.
  end if
  Receive data X_t.
  Make a selection decision $S_t \coloneqq \mathbf{S}_t(X_t)$.
  if $S_t = 1$ **then**
    $\mathcal{S}_t \coloneqq \mathcal{S}_{t-1} \cup \{t\}$.
    Construct $C_t(X_t, \alpha_t)$ for θ_t.
  else
    $\mathcal{S}_t \coloneqq \mathcal{S}_{t-1}$
  end if
end for

the online selective inference problem, e-LOND-CI and Ue-LOND-CI, respectively, that is free of both restrictons.

To ensure FCR control, both algorithms do require each C_t to a special type of CI: an *e-CI* (Vovk and Wang, 2023; Xu et al., 2022) — similar to how e-LOND applies to e-values. $C(X, \alpha)$ is an e-CI over the universe of parameters Θ if it can be written as follows:

$$C(X, \alpha) = \{\theta \in \Theta : E_\theta < \alpha^{-1}\}, \tag{9}$$

where E_θ is an e-value when the true parameter is θ. Note that the e-CI in (9) does satisfy the CI definition in (8) by Markov's inequality applied to E_{θ^*}, where θ^* is the true parameter. Let $(\mathcal{S}_t^{\text{e-LOND}})$ and $(\mathcal{S}^{\text{Ue-LOND}})$ denote the resulting selection sets of e-LOND-CI and Ue-LOND-CI, respectively. We now present our fourth main result, whose proof is in Section 8.3.

Theorem 4. *For any dependence structure among the data, (X_t), and sequence of selection rules $(\mathbf{S}_t)$,* $\text{FCR}(\mathcal{S}_t^{\text{e-LOND}}), \text{FCR}(\mathcal{S}_t^{\text{Ue-LOND}}) \leq \alpha$ *for all $t \in \mathbb{N}$.*

Remark 1. *Unlike discovery sets $(\mathcal{R}_t)$ in the the online FDR control problem, the selection sets $(\mathcal{S}_t)$* do *not depend on (α_t) — $(\mathcal{S}_t)$ can be chosen in an arbitrary fashion based on the observed data. Thus, algorithms with online FDR control do not necessarily provide provide FCR control. However, the reverse is true — FCR control implies FDR control (Weinstein and Ramdas, 2020, Section 5.2).*

As discussed by Xu et al. (2022), many existing canonical CIs are e-CIs, in the same way that many p-values are implicitly inverted e-values. This gives e-LOND-CI and Ue-LOND-CI broad applicability and utility as a default online selective inference method that is robust to the unknown dependence and arbitrary user choice of selection rule.

5 Numerical simulations

To highlight the practical behavior of our methods, we conduct two simulations, with different dependence structures, where we test the null hypothesis $H_0 : \mu \leq 0$, where μ is the mean of a distribution with support bounded in $[-4, 4]$. The first simulation is with local dependence between hypotheses, and the second is with sampling without replacement (WoR) dependence between hypotheses. In both instances, it we sample data sequentially, and hence our experiments exemplify the practicality of our new e-value based methods for the doubly sequential framework described in Section 2. In addition to simulations, we also describe an application of our methods to *online model-free selective inference under covariate shift* in Section 6, and compare the performance of our methods on real data from a protein prediction task from Jin and Candès (2023).

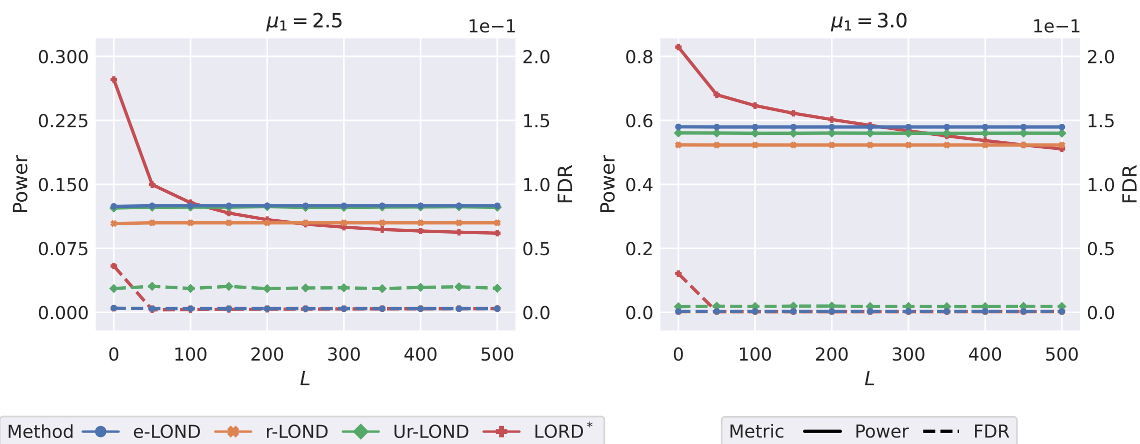

Figure 2: The power of different methods with provable FDR control against the lag parameter L in a simulation with local dependence between statistics. Empirically, the FDR of all methods is well below the desired level of $\alpha = 0.3$. As L increases (i.e., more hypotheses are dependent), we can see the power of LORD* decrease, since it is essentially ignoring hypotheses with statistics that are dependent with the current hypothesis being tested. e-LOND has consistently higher power than both p-value procedures, r-LOND and Ur-LOND, and has higher power than LORD* as L becomes large. We omit Ue-LOND since its power increase over e-LOND is very small. All Monte Carlo error from simulations is negligible (smaller than the line width in the plot).

5.1 Local dependence

We perform numerical simulations comparing e-LOND to other methods in a version of the local dependence setting from Zrnic et al. (2021). Here, we draw data in a sequential setting with bounded random variables, since powerful sequential p-values for testing the mean of bounded random variables are naturally derived from e-values. We let L be our local dependence lag parameter, i.e., the data for the tth hypothesis is independent of data from hypotheses that are more than L indices away. We let the total number of hypotheses be $T = 10^3$. For the tth hypothesis, we consider a setup where we recieve stream of $N = 200$ samples $(X_t^i)_{i \in [N]}$, where X_t^i for each $i \in [N]$ are sampled i.i.d. from a Beta distribution (shifted and scaled to be on $[\pm 4]$) with mean $\mu_0 = 0$ under the null, and $\mu_1 \in \{2.5, 3\}$ otherwise. For each $i \in [N], t \in [T]$, X_t^i has Gaussian copula dependence with $(X_{t-L}^i, \dots, X_{t+L}^i)$, i.e, the ith sample of data for hypotheses that are within L steps. Explicitly, the covariance matrix of the Gaussian distribution, Σ, is set to $\Sigma_{i,j} = 0.5^{|i-j|}$ when $|i - j| \leq L$ and 0 elsewhere. We construct p-values and e-values that are valid for this setting based on Hoeffding's inequality (see Appendix B.2 for details).

Our results are averaged over 500 trials and shown in Figure 2. In addition to comparing to r-LOND and Ur-LOND, we compare to LORD*, which is online FDR control algorithm from Zrnic et al. (2021) requires knowing the lag parameter L beforehand, so it can solely utilize test statistics from hypotheses that are independent from the current hypothesis (see Appendix B.1 for details). The power of LORD* degrades as the lag parameter increases, which is expected, since it has access to a decreasing number of discoveries. e-LOND is more powerful than both r-LOND and Ur-LOND across the board, and LORD* once $L \geq 250$ ($\mu_1 = 3$) or $L \geq 150$ ($\mu_1 = 2$). Ue-LOND only offers a small increase in power over e-LOND here so it is omitted.

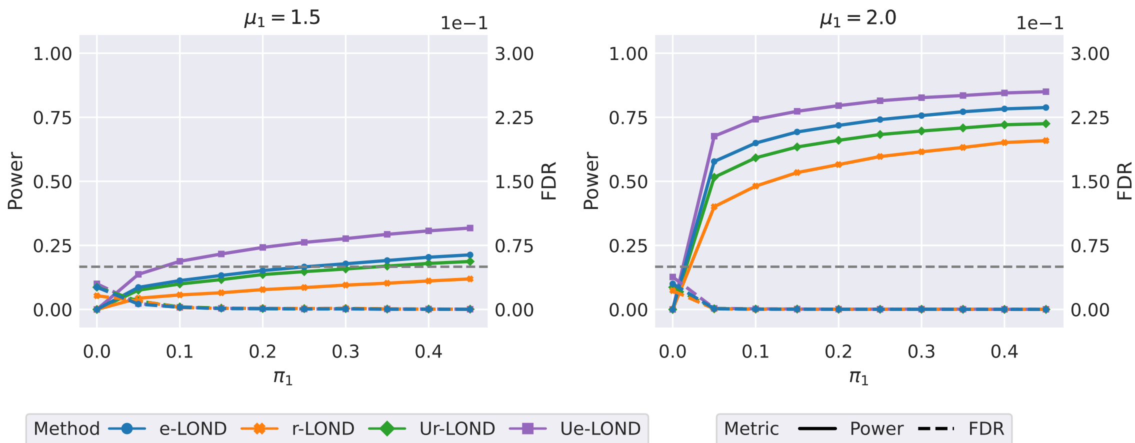

Figure 3: The power of different methods with provable FDR control against proportion of non-nulls π_1 in a simulation with sampling without replacement (WoR) dependence between statistics. Empirically, the FDR of all methods are below $\alpha = 0.05$. e-LOND has consistently higher power than both p-value procedures, r-LOND and Ur-LOND, and Ue-LOND is consistenly more powerful than e-LOND. This makes two e-value procedures, the most powerful methods. All Monte Carlo error from simulations is negligible (smaller than the line width in the plot).

5.2 Sampling WoR

We construct a population such that the mean is $\mu_0 = 0$ for the data we sample WoR for the null hypotheses, positive μ_1 for the non-null hypotheses. We will construct this population by discretizing a scaled and shifted Beta distribution. Let $V(0), V(1) \in [\pm 4]^{N \times T}$ be the populations created from $P(\mu_0), P(\mu_1)$. Let $V_{i,t}$ be the tth value in $V(i)$. We set $s = 0.01, \mu_0 = 0, \mu_1 \in \{1.5, 2\}$ in our simulations. For each simulation trial, we choose a non-null proportion $\pi_1 \in [0.1, 0.9]$, and uniformly randomly choose $B \in \{0, 1\}^T$ with exactly $\lceil \pi_1 T \rceil$ ones. Let σ be a random permutation over $[N \times T]$. Our data for the tth hypothesis is $X_t = (V_{B_t, \sigma((t-1)\cdot N + i)})_{i \in [N]}$. X_t is a sample WoR of size N from $V(0)$ if $B_t = 0$ and $V(1)$ if $B_t = 1$. Our e-values and p-values usiang an e-process for sampling WoR from Waudby-Smith and Ramdas (2020) — see Appendix B.3 for details.

Our results, averaged over 500 trials, are in Section 5.1. Here, both e-LOND and Ue-LOND dominate in power across the board, while all methods have FDR below $\alpha = 0.05$. Clearly, the theoretical improvements of our novel e-value methods translate to empirical gains.

6 Application: online model-free selective inference under covariate shift

As an application of our framework, we can address an online version of the model-free selective inference under covariate shift problem introduced by Jin and Candès (2023). To do so, we use e-LOND to directly derive an online version of the *weighted conformal selection* (WCS) procedure. In this setting, we consider labeled pairs $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$. We are given an i.i.d. calibration dataset of labeled pairs $\{(X_i, Y_i)\}_{i\in[n]}$ where $(X_i, Y_i) \sim \mathbf{P}$. Our goal is to perform inference on a stream of i.i.d. test data points $(X_{n+1}, Y_{n+1}), (X_{n+2}, Y_{n+2}), \ldots$. For each $t \in \mathbb{N}$, we only observe the covariates of the test points, X_{n+t}, and a potentially random threshold, c_{n+t}. Our goal is to test the following hypothesis about Y_{n+t}:

$$H_0^t : Y_{n+t} \leq c_{n+t}.$$

One notable difference between the setup here and standard online multiple testing is that the null hypotheses themselves are random, as Y_{n+t} and c_{n+t} are both random. However, our goal remains the same: ensure $\mathrm{FDR}(\mathcal{R}_t) \leq \alpha$ for each $t \in \mathbb{N}$ where the expectation is now also taken over the randomness of whether a hypothesis is null or not. As argued in Jin and Candès (2023), this type of selection occurs widely in practice, e.g., screening for high performing job candidates based on interview performance, picking patients with attributes that are responsive to treatment, detecting outliers, etc. In this setting with randomized null hypotheses, we require our p-values and e-values to satisfy the following conditions instead for each $t \in \mathbb{N}$:

$$\mathbb{P}\left(P_t \leq \alpha, Y_{n+t} \leq c_{n+t}\right) \leq \alpha \text{ for all } \alpha \in [0, 1], \tag{10}$$

$$\mathbb{E}[E_t \cdot \mathbf{1}\left\{Y_{n+t} \leq c_{n+t}\right\}] \leq 1. \tag{11}$$

In addition, $\mathbf{Q}$ results from a covariate shift on $\mathbf{P}$. This means that $\mathbf{P}(Y \mid X = x) = \mathbf{Q}(Y \mid X = x)$ for all $x \in \mathcal{X}$. Further, the Radon-Nikodym derivative (w.r.t. to an arbitrary common base measure) satisfies $(d\mathbf{Q}/d\mathbf{P})(x, y) = w(x)$ for all $x \in \mathcal{X}$, where w is a likelihood ratio dependent only on $x \in \mathcal{X}$. We assume we have access to w (e.g., we can esimate it from other data accurately). In addition, define a *monotone* score function $V : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ as a function satisfying $V(x, y) \leq V(x, y')$ for all $x \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$ where $y \leq y'$.

6.1 FDR control through online multiple testing

Jin and Candès (2023) construct the following p-value using any monotone score function V:

$$\begin{aligned} V_i &\coloneqq V(X_i, Y_i), \qquad \widehat{V}_{n+t} \coloneqq V(X_{n+t}, c_{n+t}), \\ P_t &\coloneqq \frac{\sum_{i=1}^n w(X_i)\mathbf{1}\{V_i < \widehat{V}_{n+t}\} + w(X_{n+t})}{\sum_{i=1}^n w(X_i) + w(X_{n+t})}, \end{aligned} \tag{12}$$

For simplicity, we assume that neither $(V_i)_{i\in[n]}$ nor $(\widehat{V}_{n+t})_{t\in\mathbb{N}}$ have point masses in their distributions in this paper, and this assumption can be relaxed through simple modifications to the p-value formulations (Jin and Candès, 2023, eqs. 3 & 6).

Fact 5 (Lemma 2.2 (Jin and Candès, 2023))**.** *For each $t \in \mathbb{N}$, P_t defined in* (12) *is a p-value* (10)*.*

The dependence structure among (P_t) is quite complicated, and does not satisfy usual independence or positive dependence notions that are amenable to multiple testing without correction (Jin and Candès, 2023, Proposition 2.4). Thus, one must apply r-LOND (or Ur-LOND) derive FDR control.

Proposition 1. *Let $(\mathcal{R}_t^{\text{r-LOND}})$ and $(\mathcal{R}_t^{\text{Ur-LOND}})$ be the sequences of rejection sets that arise from applying* r-LOND *or* Ur-LOND*, respectively, to (P_t) as defined in* (12)*. Then,* $\mathrm{FDR}(\mathcal{R}_t^{\text{r-LOND}}) \leq \alpha$ *and* $\mathrm{FDR}(\mathcal{R}_t^{\text{Ur-LOND}}) \leq \alpha$ *for each $t \in \mathbb{N}$.*

We defer the proof of this result to Section 8.4. Jin and Candès (2023) show that the more powerful way to utilize P_t is to view them as e-values, and we show that a similar phenomenon is also possible for online WCS. First, define the following leave-one-out conformal p-values $P_j^{(t),-}, P_j^{(t),+}$ for each

$t \in \mathbb{N}$ and $j \in [t-1]$:

$$P_j^{(t),-} := \frac{\sum_{i=1}^n w(X_i)\mathbf{1}\{V_i < \widehat{V}_{n+j}\}}{\sum_{i=1}^n w(X_i) + w(X_{n+t})},$$
$$P_j^{(t),+} := \frac{\sum_{i=1}^n w(X_i)\mathbf{1}\{V_i < \widehat{V}_{n+j}\} + w(X_{n+t})}{\sum_{i=1}^n w(X_i) + w(X_{n+t})}.$$

Let $\widehat{\mathcal{R}}_{t-1}^{\mathrm{LOND}(t),-}$ and $\widehat{\mathcal{R}}_{t-1}^{\mathrm{LOND}(t),+}$ be the discovery set obtained from applying LOND to $(P_j^{(t),-})_{j\in[t-1]}$ and $(P_j^{(t),+})_{j\in[t-1]}$, respectively. Define the test levels for the next hypothesis as

$$\widehat{\alpha}_t^{\mathrm{LOND},-} := \alpha\gamma_t \cdot (|\widehat{\mathcal{R}}_{t-1}^{(t),-}| + 1), \qquad \widehat{\alpha}_t^{\mathrm{LOND},+} := \alpha\gamma_t \cdot (|\widehat{\mathcal{R}}_{t-1}^{(t),+}| + 1).$$

We can now define the following e-value:

$$E_t^{\mathrm{LOND}} := \mathbf{1}\{P_t \leq \widehat{\alpha}_t^{\mathrm{LOND},+}\}/\widehat{\alpha}_t^{\mathrm{LOND},-}.$$

Proposition 2. *For each $t \in \mathbb{N}$, E_t^{LOND} is an e-value* (11)*.*

We defer the proof of this result to Section 8.5. We can derive the FDR control of e-LOND or Ue-LOND applied to (E_t^{LOND}).

Theorem 5. *Using e-values (E_t) satisfying* (11)*, $\mathrm{FDR}(\mathcal{R}_t^{\text{e-LOND}}) \leq \alpha$ and $\mathrm{FDR}(\mathcal{R}_t^{\text{Ue-LOND}}) \leq \alpha$ for each $t \in \mathbb{N}$.*

We defer the proof of this result to Section 8.6. Now, we apply our online WCS techniques to some real data settings in Jin and Candès (2023), and use their code to calculate the weighted p-values in (12) for each setup.

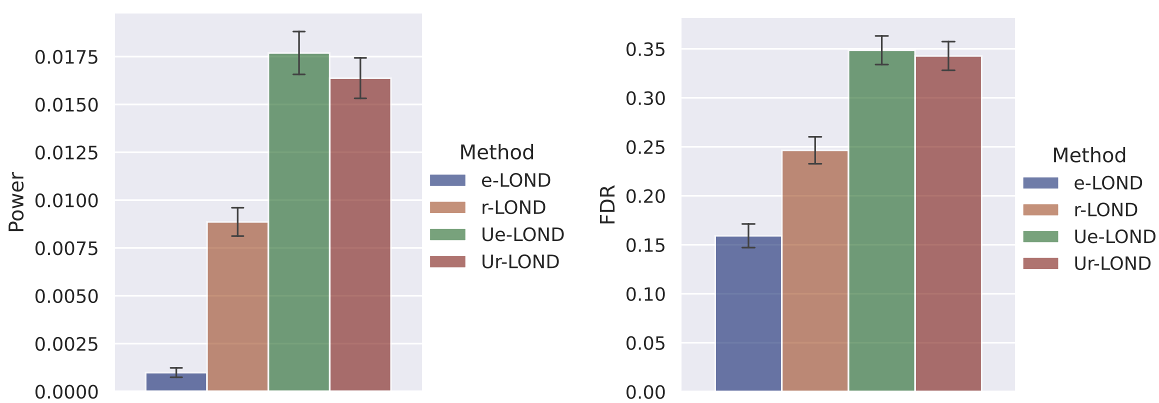

Figure 4: Average power and empirical FDR for methods applied at $\alpha = 0.5$ for the drug property prediction task, with the error bars marking one standard error from Monte Carlo estimation. We can see that Ue-LOND has the largest power. E-LOND has the smallest power, due to the way (E_t^{LOND}) are formulated to almost always be below the test threshold of e-LOND itself. However, by allowing randomization in Ue-LOND, we see that this issue is fixed and Ue-LOND exceeds the power of both Ur-LOND and r-LOND.

6.2 Drug property prediction

We tackle the task of predicting drug properties that uses the HIV screening dataset in the DeepPurpose library (Huang et al., 2021) — the goal is to select a subset of drug candidates that bind to a target protein for HIV. The covariate X is the chemical structure of the drug, that is encoded into the form of a vector $\mathbb{R}^d$, and $Y \in \{0,1\}$ is binary label of whether it does not or does bind. In constructing the calibration set, experimenters might pick drugs that seem more likely to bind to analyze (and label) and induce a covariate shift as a result of selection bias. Thus, we construct a setup that emulates this issue. 40% of the data set is placed in $\mathcal{D}_{\text{train}}$ and used to train a neural network classifier $\widehat{\mu} : \mathcal{X} \mapsto \mathbb{R}$ that predicts the probability of binding. 60% more of the dataset is used to construct $\mathcal{D}_{\text{calib}}$ by selecting each point (X_i, Y_i) to be in $\mathcal{D}_{\text{calib}}$ with probability $p(X_i)$ where $p(x) = \sigma(\widehat{\mu}(x) - \bar{\mu}) \wedge 0.8$. Of the points that are neither in $\mathcal{D}_{\text{train}}$ nor $\mathcal{D}_{\text{calib}}$, we sample 5% randomly to constitute $\mathcal{D}_{\text{test}}$ due to computational constraints. Consequently, there is a covariate shift between the calibration and the test set, and the resulting likelihood ratio satisfies $w(x) \propto 1/p(x)$. The null hypothesis that we wish to test is as follows:

$$H_0^t : Y_{n+t} = 0.$$

Controlling the FDR results in selecting a subset of drugs where only a small proportion do not bind to the protein in expectation. We average our results over 600 trials. We see that the power of Ue-LOND in Section 6.1 is the largest. On the other hand, the power of e-LOND is the smallest. This is because E_t^{LOND} is either 0 or $1/\widehat{\alpha}_t^{\text{LOND},+}$, and $\alpha_t^{\text{e-LOND}} < \widehat{\alpha}_t^{\text{LOND},+}$ holds often, as $\widehat{\alpha}_t^{\text{LOND},+}$ is a conservative estimate of $\alpha_t^{\text{e-LOND}}$. The randomization from Ue-LOND alleviates this problem, hence it attaining the largest power. All methods also practically control FDR at the desired level of $\alpha = 0.5$.

7 Related work

This work lies at the intersection of e-values and online multiple testing. We outline the most relevant research in each of these areas to this work.

Online multiple testing Online multiple testing was first posed by Foster and Stine (2008) when they were studying computationally cheap methods for performing streamed variable selection in high dimensional things and proved mFDR control for alpha-investing. The methods were subsequently improved in several follow up works to be more powerful and also guarantee control of the FDR Aharoni and Rosset (2014); Javanmard and Montanari (2018); Ramdas et al. (2017). Ramdas et al. (2018) and Tian and Ramdas (2019) developed adaptive online multiple testing procedures based on Storey's method (Storey, 2002) for offline FDR control. With the exception of Javanmard and Montanari (2018), all these works all focus on online FDR or mFDR control under the assumption that p-values are independent or are p-values when conditioned on the information observed so far (e.g., previous p-values, rejection decisions, etc.), i.e., conditional superuniformity. As mentioned before, more recent work of Zrnic et al. (2021) considers explicitly modeling dependence relationships through conflict sets to derive algorithms that still control mFDR and FDR even when independence or conditional superuniformity is not satisfied. Another line of work considers the situation when the rejection decision of a hypothesis does not have to be made immediately, but only need to be made by a later time, such as at the end of a batch of hypotheses being jointly experimented on (Zrnic et al., 2020) or at individual future deadlines (Fisher, 2022b). This is the first work that directly targets the arbitrary dependence case. Robertson et al. (2022a) provide a R package implementing many of the aforementioned methods for online control of the FDR, in addition to other online multiple testing methods. Online multiple testing methods (including LOND) have already been applied in a variety of medicinal and biological applications (Robertson and Wason, 2018; Robertson et al., 2022b; Liou et al., 2023).

E-values E-values have been applied in many offline multiple testing settings such as FDR control (Wang and Ramdas, 2022; Ignatiadis et al., 2023) and closed testing (Vovk and Wang, 2023). In particular, the e-BH procedure introduced by Wang and Ramdas (2022) has been used as a subroutine in other multiple testing procedures with FDR control such as in the bandit setting (Xu et al., 2021), for the purpose of derandomizing knockoffs (Ren and Barber, 2023) or achieving optimality under a Bayesian linear model alternative (Ahn et al., 2022). Xu et al. (2022) present selective inference procedure with FCR control for e-CIs. Further, Jin and Candès (2023) showed that the weighted conformal selection procedure in their paper can also be viewed as an application of e-BH to e-values. This work is novel in bringing all these insights concerning e-values that have been used in offline multiple testing to the online setting.

8 Omitted proofs

Here, we include the full proofs of the results contained in Section 3, Section 4, and Section 6.

8.1 Proof of Theorem 1

For brevity, we will write $\alpha_t^{\text{e-LOND}}$ as α_t in the proofs in this section.

$$\begin{aligned}
\mathrm{FDR}(\mathcal{R}_t) &= \mathbb{E}\left[\sum_{i \in \mathcal{H}_0 \cap [t]} \frac{\mathbf{1}\left\{E_i \geq \alpha_i^{-1}\right\}}{|\mathcal{R}_t| \vee 1}\right] = \sum_{i \in \mathcal{H}_0 \cap [t]} \mathbb{E}\left[\frac{\mathbf{1}\left\{E_i \geq \alpha_i^{-1}\right\}}{|\mathcal{R}_t| \vee 1} \times \mathbf{1}\left\{E_i \geq \alpha_i^{-1}\right\}\right] \\
&\overset{\text{(i)}}{\leq} \sum_{i \in \mathcal{H}_0 \cap [t]} \mathbb{E}\left[\frac{\alpha_i E_i}{|\mathcal{R}_t| \vee 1} \times \mathbf{1}\left\{|\mathcal{R}_t| \geq |\mathcal{R}_{i-1}| + 1\right\}\right] \\
&\overset{\text{(ii)}}{\leq} \sum_{i \in \mathcal{H}_0 \cap [t]} \mathbb{E}\left[\frac{\alpha\gamma_i(|\mathcal{R}_{i-1}| + 1)E_i}{|\mathcal{R}_{i-1}| + 1} \times \mathbf{1}\left\{|\mathcal{R}_t| \geq |\mathcal{R}_{i-1}| + 1\right\}\right] \\
&\overset{\text{(iii)}}{\leq} \sum_{i \in \mathcal{H}_0 \cap [t]} \mathbb{E}\left[\frac{\alpha\gamma_i(|\mathcal{R}_{i-1}| + 1)E_i}{|\mathcal{R}_{i-1}| + 1}\right] = \alpha \sum_{i \in \mathcal{H}_0 \cap [t]} \gamma_i \mathbb{E}\left[E_i\right] \leq \alpha.
\end{aligned}$$

Inequality (i) is a result of (6) and $|\mathcal{R}_t| \geq |\mathcal{R}_{i-1}| + \mathbf{1}\left\{E_i \geq \alpha_i^{-1}\right\}$ by construction of $\mathcal{R}_t$. Inequality (i) is a result of the indicator in the expectation (i.e., making discovery at H_i will make $\mathcal{R}_t$ larger than $\mathcal{R}_{i-1}$). Inequality (iii) comes from dropping the indicator term. The last inequality is due to $\mathbb{E}[E_t] \leq 1$ for all $t \in \mathcal{H}_0$ by definition of e-values (3), and because (γ_t) sum up to 1. Thus, we achieve an upper bound of α on the final line and have shown our desired result on FDR.

To show e-LOND strictly dominates r-LOND, it is sufficient show that $\alpha_t^{\text{e-LOND}} \geq \alpha_t^{\text{r-LOND}}$ for all $t \in \mathbb{N}$, and there exists a sequence of e-values (E_t) such that there exists $t \in \mathbb{N}$ such that $\alpha_t^{\text{e-LOND}} > \alpha_t^{\text{r-LOND}}$. For any $t \in \mathbb{N}$,

$$\beta_t(|\mathcal{R}_{t-1}| + 1) = \int_0^{|\mathcal{R}_{t-1}|+1} x \; d\nu(x) \leq |\mathcal{R}_{t-1}| + 1,$$

where the first equality is by definition of reshaping function, and the inequality is because $x \leq |\mathcal{R}_{t-1}|+1$ in the integrand, and ν is a probability measure that is nonnegative and integrates to 1. Thus, $\alpha_t^{\text{e-LOND}} \geq \alpha_t^{\text{r-LOND}}$ for all $t \in \mathbb{N}$.

Next, note for β_2, either it satisfes (1) $\beta_2(2) = 2$ and $\beta_2(1) = 0$ or (2) $\beta_2(2) < 2$ — this follows from the definition of reshaping function, and case (1) correpsonds to putting all probability mass in ν on 2.

If β_2 satisfies case (1), then we set $E_1 = 1/(\alpha\gamma_1)+1$. This results in $\alpha_2^{\text{e-LOND}} = \alpha\gamma_2 > 0 = \alpha_2^{\text{r-LOND}}$. Otherwise, we set $E_1 = 1/(\alpha\gamma_1)$, which leads to a rejection by e-LOND, and note that $\alpha_2^{\text{r-LOND}} \leq \alpha\gamma_2\beta_2(2) < 2\alpha\gamma_2 = \alpha_2^{\text{e-LOND}}$. Thus, we have shown that e-LOND strictly dominates r-LOND applied to $(1/E_t)$ and conclude our proof. □

8.2 Proof of Theorem 3

For simplicity, denote $\alpha_t^{\text{Ur-LOND}}, \mathcal{R}_t^{\text{Ur-LOND}}$ as $\alpha_t, \mathcal{R}_t$. Similar to the proof of FDR control for r-LOND in Zrnic et al. (2021), we first show the following inequality for any $i \in [t]$:

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \alpha_i^{\text{Ur-LOND}}\right\}}{|\mathcal{R}_t| \vee 1}\right] &\overset{\text{(i)}}{=} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \alpha_i^{\text{Ur-LOND}}\right\}}{|\mathcal{R}_t| \vee 1}\mathbf{1}\left\{|\mathcal{R}_{t-1}| \geq |\mathcal{R}_{i-1}| + 1\right\}\right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \alpha\gamma_i\beta_i((|\mathcal{R}_{i-1}|+1)/U_i)\right\}}{|\mathcal{R}_t| \vee 1}\mathbf{1}\left\{|\mathcal{R}_t| \vee 1 \geq |\mathcal{R}_{i-1}| + 1\right\}\right] \\
&\overset{\text{(iii)}}{\leq} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \alpha\gamma_i\beta_i((|\mathcal{R}_{i-1}|+1)/U_i)\right\}}{|\mathcal{R}_{i-1}| + 1}\mathbf{1}\left\{|\mathcal{R}_t| \vee 1 \geq |\mathcal{R}_{i-1}| + 1\right\}\right] \\
&\overset{\text{(iv)}}{\leq} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \alpha\gamma_i\beta_i((|\mathcal{R}_{i-1}|+1)/U_i)\right\}}{|\mathcal{R}_{i-1}| + 1}\right] \overset{\text{(v)}}{\leq} \alpha\gamma_t.
\end{aligned} \tag{13}$$

Equality (i) is because $\{P_i \leq \alpha_i^{\text{Ur-LOND}}\} \Rightarrow \{|\mathcal{R}_{i-1}| + 1 \leq |\mathcal{R}_t| \vee 1\}$ as a result of a discovery being made at the ith hypothesis. Equality (ii) is by expanding the definition of $\alpha_i^{\text{Ur-LOND}}$. Inequality (iii) is the indicator $\mathbf{1}\{|\mathcal{R}_t| \vee 1 \geq |\mathcal{R}_{i-1}| + 1\}$ being 1 iff the event it is indicating is true. Inequality (iv) is simply by droppign the indicator. Inequality (v) is by Fact 4. Thus, we can derive the following bound on the FDR by (13):

$$\begin{aligned}
\mathrm{FDR}(\mathcal{R}_t) &= \sum_{i \in \mathcal{H}_0 \cap [t]} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \alpha_i^{\text{Ur-LOND}}\right\}}{|\mathcal{R}_t| \vee 1}\right] \\
&\leq \alpha \sum_{[t]} \gamma_t \leq \alpha,
\end{aligned}$$

which achieves our desired FDR control.

The strict dominance in expectation follows from the fact that $\alpha_t^{\text{Ur-LOND}} > \alpha_t^{\text{r-LOND}}$ with nonzero probability whenever $|\mathcal{R}_{t-1}| < t - 1$ because U_t^{-1} is a positive number that is at least 1, and $(|\mathcal{R}_{t-1}| + 1)U_t^{-1} \geq |\mathcal{R}_{t-1}| + 2$ (which implies $\beta_t^{\text{BY}}((|\mathcal{R}_{t-1}| + 1)U_t^{-1}) > \beta_t^{\text{BY}}(|\mathcal{R}_{t-1}| + 1)$) with nonzero probability. Thus, we have shown strict dominance in expectation and all results in the theorem. □

8.3 Proof of Theorem 4

Denote $\alpha_t^{\text{e-LOND}}$ as α_t in this section. We make the following derivation for the FCR:

$$\begin{aligned}
\mathrm{FCR}(\mathcal{S}_t) = \mathbb{E}\left[\sum_{i \in \mathcal{S}_t} \frac{\mathbf{1}\{\theta_i \not\in C_i(X_i, \alpha_i)\}}{|\mathcal{S}_t| \vee 1}\right] &\overset{\text{(i)}}{=} \mathbb{E}\left[\sum_{i \in \mathcal{S}_t} \frac{\mathbf{1}\left\{E_{\theta_i} \geq \alpha_i^{-1}\right\}}{|\mathcal{S}_t| \vee 1}\right] \\
\overset{\text{(ii)}}{\leq} \mathbb{E}\left[\sum_{i \in \mathcal{S}_t} \frac{\alpha\gamma_i(|\mathcal{S}_{i-1}| + 1)E_i}{|\mathcal{S}_t| \vee 1}\right] &\overset{\text{(iii)}}{=} \sum_{i \in [t]} \mathbb{E}\left[\frac{\alpha\gamma_i(|\mathcal{S}_{i-1}| + 1)E_i\mathbf{1}\{i \in \mathcal{S}_t\}}{|\mathcal{S}_t| \vee 1}\right] \\
\overset{\text{(iv)}}{\leq} \sum_{i \in [t]} \mathbb{E}\left[\frac{\alpha\gamma_i(|\mathcal{S}_{i-1}| + 1)E_{\theta_i}\mathbf{1}\{|\mathcal{S}_t| \geq |\mathcal{S}_{i-1}| + 1\}}{|\mathcal{S}_t| \vee 1}\right]& \\
\overset{\text{(v)}}{\leq} \sum_{i \in [t]} \mathbb{E}\left[\frac{\alpha\gamma_i(|\mathcal{S}_{i-1}| + 1)E_{\theta_i}\mathbf{1}\{|\mathcal{S}_t| \vee 1 \geq |\mathcal{S}_{i-1}| + 1\}}{|\mathcal{S}_{i-1}| + 1}\right]& \\
\overset{\text{(vi)}}{\leq} \sum_{i \in [t]} \mathbb{E}\left[\frac{\alpha\gamma_i(|\mathcal{S}_{i-1}| + 1)E_{\theta_i}}{|\mathcal{S}_{i-1}| + 1}\right] = \alpha \sum_{i \in [t]} \gamma_i \mathbb{E}\left[E_{\theta_i}\right] \leq \alpha.&
\end{aligned}$$

Equality (i) is by the definition of an e-CI in (9). Inequality (ii) is by the definition of an e-LOND. Equality (iii) is simply arithmetic with the indicator of whether i is in $\mathcal{S}_t$. Inequality (iv) is because $i \in \mathcal{S}_t$ implies that $\mathcal{S}_t$ gained a selected parameter, namely the ith parameter, over $\mathcal{S}_{i-1}$. Inequality (v) is because $i \in \mathcal{S}_t$ implies that $\mathcal{S}_t$ gained a selected parameter, namely the ith parameter, over $\mathcal{S}_{i-1}$. Inequality (vi) follows from dropping the indicator, and the last inequality is again due to $\mathbb{E}[E_{\theta_i}] \leq 1$ for each $i \in \mathbb{N}$ by definition of e-values (3), and because (γ_t) sum up to 1. Thus, we achieve our desired result of FCR control of α.

Ue-LOND-CI can be shown to have FCR control by following the above argument, except we can replace E_{θ_i} with $S_{\alpha_t^{\text{e-LOND}}}(E_{\theta_i})$. Thus, we have shown our desired levels of FCR control. □

8.4 Proof of Proposition 1

Let $\widetilde{P}_t \coloneqq P_t \vee \mathbf{1}\left\{Y_t > c_t\right\}$. Note that for each $t \in \mathbb{N}$, $\widetilde{P}_t$ satisfies the following two properties.

$$\{\widetilde{P}_t \leq s\} \Leftrightarrow \{P_t \leq s, Y_t \leq c_t\} \text{ and } \mathbb{P}\left(\widetilde{P}_t \leq s\right) = \mathbb{P}\left(P_t \leq s, Y_t \leq c_t\right) \leq s \text{ for all } s \in [0, 1). \tag{14}$$

This is by definition of $\widetilde{P}_t$ and by the superuniform constraint on P_t in (10). Further, we can see that

$$\{P_t \leq s, Y_t \leq c_t\} \Rightarrow \{\widetilde{P}_t \leq s\} \text{ when } s = 1, \tag{15}$$

by definition of $\widetilde{P}_t$ as well.

We also observe the following implication holds:

$$\{\widetilde{P}_i \leq \alpha_i\} \Rightarrow \{\mathcal{R}_t \supset \mathcal{R}_{i-1}\} \Rightarrow \{|\mathcal{R}_t| \vee 1 \geq \mathcal{R}_{i-1} + 1\}, \tag{16}$$

for all $t \geq i$ simply because a discovery set grows when a a new discovery is made.

Let $(\alpha_t), (\mathcal{R}_t)$ be either $(\alpha_t^{\text{r-LOND}}), (\mathcal{R}_t^{\text{r-LOND}})$ or $(\alpha_t^{\text{Ur-LOND}}), (\mathcal{R}_t^{\text{Ur-LOND}})$. We can make the following derivation of the FDR:

$$\begin{aligned}
\mathrm{FDR}(\mathcal{R}_t) &= \sum_{i \in [t]} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \alpha_t, i \in \mathcal{H}_0\right\}}{|\mathcal{R}_t| \vee 1}\right] = \sum_{i \in [t]} \mathbb{E}\left[\frac{\mathbf{1}\left\{P_i \leq \alpha_t, Y_i \leq c_i\right\}}{|\mathcal{R}_t| \vee 1}\right] \overset{\text{(i)}}{\leq} \sum_{i \in [t]} \mathbb{E}\left[\frac{\mathbf{1}\left\{\widetilde{P}_i \leq \alpha_i\right\}}{|\mathcal{R}_t| \vee 1}\right] \\
&\overset{\text{(ii)}}{=} \sum_{i \in [t]} \mathbb{E}\left[\frac{\mathbf{1}\left\{\widetilde{P}_i \leq \alpha_i\right\}}{|\mathcal{R}_{i-1}| + 1}\right] \overset{\text{(iii)}}{\leq} \sum_{i \in [t]} \mathbb{E}\left[\frac{\mathbf{1}\left\{\widetilde{P}_i \leq \alpha\gamma_i \cdot \beta_i((|\mathcal{R}_{i-1}| + 1)/U_i)\right\}}{|\mathcal{R}_{i-1}| + 1}\right] \overset{\text{(iv)}}{\leq} \sum_{i \in [t]} \alpha\gamma_i \leq \alpha.
\end{aligned}$$

Inequality (i) is by a combination of (14) and (15). Inequality (ii) is because of (16). Inequality (iii) is by the definition of either choice of (α_t) ($U_i = 1$ if r-LOND, and U_i is an independent uniform random variable over $[0, 1]$ if Ur-LOND) and the fact that $|\mathcal{R}_t| \vee 1 \leq |\mathcal{R}_{t-1}| + 1$ by definition of discovery sets. Inequality (iv) is by Fact 4, since U_i is superuniform and independent of all $\widetilde{P}_i$. The last inequality is due to $\sum_{i \in [t]} \gamma_i \leq 1$. Thus, we have shown our desired FDR control. □

8.5 Proof of Proposition 2

We follow a similar proof structure to the proof of Theorem 3.1 in Jin and Candès (2023).

First, we define the following oracle p-values (that cannot be computed from the observable data) to assist with our proof:

$$\begin{aligned}
\bar{P}_t &\coloneqq \frac{\sum_{i=1}^n w(X_i)\mathbf{1}\{V_i < V_{n+t}\} + w(X_{n+t})}{\sum_{i=1}^n w(X_i) + w(X_{n+t})}. \\
\bar{P}_j^{(t)} &\coloneqq \frac{\sum_{i=1}^n w(X_i)\mathbf{1}\{V_i < \widehat{V}_{n+j}\} + w(X_{n+t})\mathbf{1}\{V_{n+t} < \widehat{V}_{n+j}\}}{\sum_{i=1}^n w(X_i) + w(X_{n+t})}.
\end{aligned}$$

These essentially replace $\widehat{V}_{n+t}$ with V_{n+t} when compared to their empirical counterparts P_t and $P_j^{(t)}$, respectively. The first thing we note is the following relationship between the oracle nonconformity score and the empirical nonconformity score at $n+t$:

$$t \in \mathcal{H}_0 \Leftrightarrow Y_{n+t} \leq c_{n+t} \Rightarrow V_{n+t} \leq \widehat{V}_{n+t} \Rightarrow \bar{P}_t \leq \widehat{P}_t,$$

since V is a monotone score function. Further, the oracle p-values $(\bar{P}_j^{(t)})_{j\in[t-1]}$ are bounded by their empirical counterparts, i.e.,

$$\widehat{P}_j^{(t),-} \leq \bar{P}_j^{(t)} \leq \widehat{P}_j^{(t),+} \text{ for all } t \in \mathbb{N} \text{ and } j \in [t-1]. \tag{17}$$

Define $\bar{\mathcal{R}}_{t-1}$ to be the discovery set that results from applying LOND to $(\bar{P}_1^{(t)}, \dots, \bar{P}_{t-1}^{(t)})$, and define

$$\bar{\alpha}_t^{\mathrm{LOND}} \coloneqq \alpha\gamma_t \cdot (|\bar{\mathcal{R}}_{t-1}| + 1), \qquad \bar{E}_t^{\mathrm{LOND}} \coloneqq \mathbf{1}\left\{\bar{P}_t \leq \bar{\alpha}_t^{\mathrm{LOND}}\right\} / \bar{\alpha}_t^{\mathrm{LOND}}$$

to be the test level for the next hypothesis and an all-or-nothing e-value testing at that level, respectively. By (17), we can derive that

$$|\widehat{\mathcal{R}}_{t-1}^{+}| \leq |\bar{\mathcal{R}}_{t-1}| \leq |\widehat{\mathcal{R}}_{t-1}^{-}|, \text{ and } \widehat{\alpha}_t^{\mathrm{LOND},+} \leq \bar{\alpha}_t^{\mathrm{LOND}} \leq \widehat{\alpha}_t^{\mathrm{LOND},-}.$$

This gives us the following inequality:

$$\mathbf{1}\{t \in \mathcal{H}_0\} \cdot E_t^{\mathrm{LOND}} = \frac{\mathbf{1}\{t \in \mathcal{H}_0\} \cdot \mathbf{1}\left\{\widehat{P}_t \leq \widehat{\alpha}_t^{\mathrm{LOND},+}\right\}}{\widehat{\alpha}_t^{\mathrm{LOND},-}} \leq \frac{\mathbf{1}\{t \in \mathcal{H}_0\} \cdot \mathbf{1}\left\{\bar{P}_t \leq \bar{\alpha}_t^{\mathrm{LOND}}\right\}}{\bar{\alpha}_t^{\mathrm{LOND}}} \leq \bar{E}_t^{\mathrm{LOND}}. \tag{18}$$

Now we need to show that $\bar{E}_t^{\mathrm{LOND}}$ is an e-value as defined in (11). Define $Z_i \coloneqq (X_i, Y_i)$ for each $i \in \mathbb{N}$. Let $Z \coloneqq [Z_1, \dots, Z_n, Z_{n+t}]$ denote the unordered set of $\{Z_1, \dots, Z_n, Z_{n+t}\}$, and $z = [z_1, \dots, z_n, z_{n+t}]$ be the unordered set of their realized values. Define $\xi_{z,t}$ as the event such that $Z = z$. Let $I_t \in [n] \cup \{n+t\}$ be the index such that $Z_{n+t} = z_{I_t}$. Now, we note the following important facts

$$\bar{P}_t \text{ is measurable w.r.t. } Z \text{ and } I_t.$$
$$(\bar{P}_j^{(t)})_{j\in[t-1]}, \bar{\mathcal{R}}_{t-1}, \bar{\alpha}_t^{\mathrm{LOND}} \text{ are measurable w.r.t. } Z \text{ and } \{Z_{n+i}\}_{i\neq t}.$$

In addition, we have that

$$\{Z_{n+i}\}_{i\neq t} \perp\!\!\!\perp I_t \mid \xi_{t,z}.$$

This is a result of $\{Z_{n+i}\}_{i\neq t} \perp\!\!\!\perp \{Z_i\}_{i\in[n]\cup\{n+t\}}$ since each data point is assumed to be independent. As a result, we can conclude that

$$\bar{P}_t \perp\!\!\!\perp \bar{\alpha}_t^{\mathrm{LOND}} \mid \xi_{z,t}. \tag{19}$$

Let $F_{z,t} \coloneqq \mathbb{P}\left(\bar{P}_t \leq \bar{\alpha}_t^{\mathrm{LOND}} \mid \xi_{z,t}\right)$ be the conditional c.d.f. of $\bar{P}_t$.

Now, we define a randomized oracle conformal p-value:

$$P_t^* \coloneqq \frac{\sum_{i=1}^n w(X_i)\mathbf{1}\{V_i < V_{n+t}\} + U_t^*(w(X_{n+t}) + \mathbf{1}\{V_i = V_{n+t}\})}{\sum_{i=1}^n w(X_i) + w(X_{n+t})}.$$

where U_t^* is an independent uniform random variable on $[0, 1]$.

We know cite the following fact from Hu and Lei (2023) that arises due to weighted exchangeability of $(Z_1, \dots, Z_n, Z_{n+t})$:

Fact 6 (Lemmas 2 and 3 of Hu and Lei (2023))**.** $P_t^* \mid \xi_{t,j}$ *is uniformly distributed over* $[0,1]$.

Since $P_t^* \leq \bar{P}_t$ determinstically, we have that

$$F_{z,t}(s) \leq \mathbb{P}\left(P_t^* \leq s \mid \xi_{z,t}\right) \leq s \text{ for all } s \in [0,1]. \tag{20}$$

Relating this back to our e-value, we, get that

$$\mathbb{E}[\bar{E}_t^{\text{LOND}} \mid \xi_{z,t}] = F_{z,t}(\bar{\alpha}_t^{\text{LOND}})/\bar{\alpha}_t^{\text{LOND}} \leq 1 \tag{21}$$

by (20) and (19). $\mathbb{E}[\bar{E}_t^{\text{LOND}}] \leq 1$ follows by the tower property of conditional expectation applied to (21). Hence, our desired result that E_t^{LOND} is an e-value follows from (18).

8.6 Proof of Theorem 5

Let $\alpha_t, \mathcal{R}_t$ be short for $\alpha_t^{\text{e-LOND}}, \mathcal{R}_t^{\text{e-LOND}}$. We can make the following derivation:

$$\begin{aligned}
\text{FDR}(\mathcal{R}_t) &= \sum_{i\in[t]} \mathbb{E}\left[\frac{\mathbf{1}\left\{E_i \geq \alpha_i, i \in \mathcal{H}_0\right\}}{|\mathcal{R}_t| \vee 1}\right] = \sum_{i\in[t]} \mathbb{E}\left[\frac{\mathbf{1}\left\{E_i \geq \alpha_i\right\} \cdot \mathbf{1}\left\{i \in \mathcal{H}_0\right\}}{|\mathcal{R}_t| \vee 1}\right] \\
&\overset{\text{(i)}}{=} \sum_{i\in[t]} \mathbb{E}\left[\frac{\mathbf{1}\left\{E_i \geq \alpha_i\right\} \cdot \mathbf{1}\left\{i \in \mathcal{H}_0\right\}}{|\mathcal{R}_t| \vee 1} \cdot \mathbf{1}\left\{|\mathcal{R}_t| \geq |\mathcal{R}_{i-1}| + 1\right\}\right] \\
&\overset{\text{(ii)}}{\leq} \sum_{i\in[t]} \mathbb{E}\left[\frac{\alpha_i E_i \cdot \mathbf{1}\left\{i \in \mathcal{H}_0\right\}}{|\mathcal{R}_t| \vee 1} \cdot \mathbf{1}\left\{|\mathcal{R}_t| \vee 1 \geq |\mathcal{R}_{i-1}| + 1\right\}\right] \\
&\overset{\text{(iii)}}{\leq} \sum_{i\in[t]} \mathbb{E}\left[\frac{\alpha_i E_i \cdot \mathbf{1}\left\{i \in \mathcal{H}_0\right\}}{|\mathcal{R}_{i-1}| + 1}\right] \overset{\text{(iv)}}{=} \sum_{i\in[t]} \mathbb{E}\left[\frac{\alpha\gamma_i(|\mathcal{R}_{i-1}| + 1) E_i \cdot \mathbf{1}\left\{i \in \mathcal{H}_0\right\}}{|\mathcal{R}_{i-1}| + 1}\right] \\
&= \sum_{i\in[t]} \alpha\gamma_i \mathbb{E}\left[E_i \cdot \mathbf{1}\left\{i \in \mathcal{H}_0\right\}\right] \leq \sum_{i\in[t]} \alpha\gamma_i \leq \alpha.
\end{aligned}$$

Inequality (i) is because $E_i \geq \alpha_i$ implies a discovery is made at the ith hypothesis. Inequality (ii) is because E_i, α_i are nonnegative. Inequality (iii) is a result of dropping the indicator for $|\mathcal{R}_t| \vee 1 \geq |\mathcal{R}_{i-1}| + 1$ and lower bounding the denominator. Equality (iv) is by exanding the definition of α_t and the final two inequalities are by the definition of an e-value from (11) and $\sum_i \gamma_i \leq 1$. FDR control of Ue-LOND can be proven in a similar fashion by replacing E_i with $S_{\alpha_i^{\text{e-LOND}}}(E_i)$, since $\mathbb{E}[S_{\alpha_i^{\text{e-LOND}}}(E_i) \mid E_i] = E_i$. Thus, we know that $S_{\alpha_i^{\text{e-LOND}}}(E_i)$ is also an e-value as defined in (11) by the tower property of conditional expectation, and the rest of the proof follows.

9 Conclusion

E-LOND and Ue-LOND are two novel procedures that utilize e-values to provide state-of-the-art performance, both practically and theoretically, in power while ensuring provable FDR control under arbitrary dependence. We also built on recent results in using randomization for multiple testing to develop the more powerful randomized online multiple testing procedures of Ue-LOND and Ur-LOND. One natural direction is to extend our results to the LORD family of algorithms, which are more powerful, but assign test levels based on the number of hypotheses between the current hypothesis and each of the previous rejections – more careful analysis is required to ensure FDR control. Note that the sharpness result in Appendix C does not preclude this possibility because it only shows that the FDR e-LOND is tight in one specific instance, but e-LOND could be improved in other instances (e.g., have larger test levels when at least one discovery is made). Current LORD algorithms rely on independence and PRDS assumptions to have FDR control while retaining power. Another direction is to explore how e-values can be incorporated with the adaptive online FDR controlling procedures of SAFFRON (Ramdas et al., 2018) and ADDIS (Tian and Ramdas, 2019), which estimate the proportion of nulls in the manner of Storey-BH (Storey, 2002).

Acknowledgements The authors acknowledge support from NSF grant DMS-1916320.

References

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571): 68–74, 2015.

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

Taejoo Ahn, Licong Lin, and Song Mei. Near-optimal multiple testing in Bayesian linear models with finite-sample FDR control. arXiv:2211.02778, 2022.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

Gilles Blanchard and Etienne Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008.

Casper Solheim Bojer and Jens Peder Meldgaard. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 37(2):587–603, 2021.

M. J. Campbell, A. Donner, and N. Klar. Developments in cluster randomized trials and Statistics in Medicine. *Statistics in Medicine*, 26(1):2–19, 2007.

Robin Dunn, Aditya Gangrade, Larry Wasserman, and Aaditya Ramdas. Universal Inference Meets Random Projections: A Scalable Test for Log-concavity. arXiv:2111.09254, 2022.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving Statistical Validity in Adaptive Data Analysis. In *ACM Symposium on Theory of Computing*, 2015.

Aaron Fisher. Online False Discovery Rate Control for LORD & SAFFRON Under Positive, Local Dependence. arXiv:2110.08161, 2022a.

Aaron J. Fisher. Online Control of the False Discovery Rate under "Decision Deadlines". In *International Conference on Artificial Intelligence and Statistics*, 2022b.

Dean Foster and Robert A Stine. Alpha-Investing: A Procedure for Sequential Control of Expected False Discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2): 429, 2008.

Aditya Gangrade, Alessandro Rinaldo, and Aaditya Ramdas. A Sequential Test for Log-Concavity. arXiv:2301.03542, 2023.

Xiaoyu Hu and Jing Lei. A Two-Sample Conditional Distribution Test Using Conformal Prediction and Weighted Rank Sum. *Journal of the American Statistical Association*, 0(0):1–19, 2023.

Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. DeepPurpose: A deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23):5545–5547, 2021.

Nikolaos Ignatiadis, Ruodu Wang, and Aaditya Ramdas. E-values as unnormalized weights in multiple testing. *Biometrika (forthcoming)*, 2023.

Adel Javanmard and Andrea Montanari. On Online Control of False Discovery Rate. arXiv:1502.06197, 2015.

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

Ying Jin and Emmanuel J. Candès. Model-free selective inference under covariate shift via weighted conformal p-values. arXiv:2307.09291, 2023.

Gautier Koscielny, Gagarine Yaikhom, Vivek Iyer, Terrence F. Meehan, Hugh Morgan, Julian Atienza-Herrero, Andrew Blake, Chao-Kung Chen, Richard Easty, Armida Di Fenza, Tanja Fiegel, Mark Grifiths, Alan Horne, Natasha A. Karp, Natalja Kurbatova, Jeremy C. Mason, Peter Matthews, Darren J. Oakley, Asfand Qazi, Jack Regnart, Ahmad Retha, Luis A. Santos, Duncan J. Sneddon, Jonathan Warren, Henrik Westerberg, Robert J. Wilson, David G. Melvin, Damian Smedley, Steve D. M. Brown, Paul Flicek, William C. Skarnes, Ann-Marie Mallon, and Helen Parkinson. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Research*, 42(D1):D802–D809, 2014.

Lathan Liou, Milena Hornburg, and David S Robertson. Global FDR control across multiple RNAseq experiments. *Bioinformatics*, 39(1), 2023.

DJ Newman, S Hettich, CL Blake, and CJ Merz. Uci repository of machine learning databases. irvine, ca: University of california, department of information and computer science. *See also http://www. ics. uci. edu/˜ mlearn/MLRepository. html*, 1998.

Aaditya Ramdas, Fanny Yang, Martin J Wainwright, and Michael I Jordan. Online control of the false discovery rate with decaying memory. In *Neural Information Processing Systems*, 2017.

Aaditya Ramdas, Tijana Zrnic, Martin Wainwright, and Michael Jordan. SAFFRON: an adaptive algorithm for online control of the false discovery rate. In *International Conference on Machine Learning*, 2018.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. arXiv:2009.03167, 2020.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M. Koolen. How can one test if a binary sequence is exchangeable? Fork-convex hulls, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 2021.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science (forthcoming)*, 2023.

Zhimei Ren and Rina Foygel Barber. Derandomised knockoffs: Leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2023.

David S Robertson and James Wason. Online control of the false discovery rate in biomedical research. arXiv:1809.07292, 2018.

David S. Robertson, Lathan Liou, Aaditya Ramdas, and Natasha A. Karp. *onlineFDR: Online error control*, 2022a. R package 2.6.0.

David S. Robertson, James M. S. Wason, Franz König, Martin Posch, and Thomas Jaki. Online error control for platform trials. arXiv:2202.03838, 2022b.

David S. Robertson, James M. S. Wason, and Aaditya Ramdas. Online multiple hypothesis testing for reproducible research. *Statistical Science (forthcoming)*, 2023.

Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472, 1990.

John David Storey. *False Discovery Rates Theory and Applications to DNA Microarrays*. PhD thesis, Stanford University, 2002.

Jinjin Tian and Aaditya Ramdas. ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls. In *Neural Information Processing Systems*, 2019.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Vladimir Vovk and Ruodu Wang. Confidence and discoveries with e-values. *Statistical Science*, 2023.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84:822–852, 2022.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *Neural Information Processing Systems*, 2020.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2023.

Asaf Weinstein and Aaditya Ramdas. Online control of the false coverage rate and false sign rate. In *International Conference on Machine Learning*, 2020.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. From infrastructure to culture: A/B testing challenges in large scale social networks. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015.

Ziyu Xu and Aaditya Ramdas. More powerful multiple testing under dependence via randomization. arXiv:2305.11126, 2023.

Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. A unified framework for bandit multiple testing. In *Neural Information Processing Systems*, 2021.

Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. Post-selection inference for e-value based confidence intervals. arXiv:2203.12572, 2022.

Fanny Yang, Aaditya Ramdas, Kevin G Jamieson, and Martin J Wainwright. A framework for Multi-A(rmed)/B(andit) Testing with Online FDR Control. In *Neural Information Processing Systems*, 2017.

Tijana Zrnic, Daniel Jiang, Aaditya Ramdas, and Michael Jordan. The Power of Batching in Multiple Hypothesis Testing. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Tijana Zrnic, Aaditya Ramdas, and Michael I. Jordan. Asynchronous Online Testing of Multiple Hypotheses. *Journal of Machine Learning Research*, 22(33):1–39, 2021.

A Comments on the online multiple testing problem

We provide additional comments on the motivation behind the formlation of the online multiple testing problem in this section by discussing why FDR is our target error metric and the relationship between online multiple testing and adaptive data analysis.

A.1 Additional remarks on online FDR control

One might wonder why we wish to simply ensure FDR control, and not prove guarantees about the power of our algorithms as well, e.g., the expected proportion of non-null hypotheses that we actually discover with our algorithm. This is because the in scientific discovery, we cannot know the exact distribution of the statistic under the true distribution when the null hypothesis is false—that would defeat the purpose of testing if the null hypothesis is true in the first place. Prior knowledge or assumptions about the distribution of the true distribution when the null hypothesis is false is often already incorporated by the scientist when designing the individual statistics that are passed to the online multiple testing algorithm. Hence, our framework for online FDR control allows for the user to flexibily change α_t to be large or small based on what they expect the signal of the hypothesis to be.

A.2 Relating online multiple testing and adaptive data analysis

There is a rich literature on *adaptive data analysis* (Dwork et al., 2015) that explicitly tackles the data reuse problem, but it is orthogonal to our setup as it focused on the problem of estimation, makes assumptions about the statistic (e.g., bounded) being tested, and focuses on the relation between the number of adaptively chosen parameters can be accurately estimated and the number of i.i.d. samples that have been gathered. On the other hand, online multiple testing is agnostic to the exact data generating mechanism (e.g., single dataset, data gathered in a correlated fashion, datasets being merged together, etc.), assumes access to the data only through a statistic (i.e., p-value, e-value, or CI), and maintains error control for a potentially infinite stream of hypotheses, which are not assumed to be adaptively or adversarially chosen. Hence, these two approaches are complementary to each other — adaptive data analysis focuses on what is the max number of parameters one can estimate for a fixed set of data, while online multiple testing aims to ensure Type I error control regardless of the underlying data sampling method used to test each hypothesis.

B Simulation details

We provide the details of our simulations (in Section 5) in this section. In this section, any references to discount sequence (γ_t) is referring to the same choice of (γ_t) used in the corresponding algorithm (i.e., e-LOND, Ue-LOND, r-LOND, or Ur-LOND) that is acting on the e-values or p-values. In all our simulations, we let $\gamma_t = 1/(t(t+1))$. We ran the simulations on a 12 core, 60GB RAM cloud server.

B.1 Definition of LORD*

We recall the LORD* algorithm of Zrnic et al. (2021) as follows:

$$\alpha_t^{\mathrm{LORD}^*} := \alpha \left(w_0 \gamma_t + \mathbf{1}\left\{ |\mathcal{R}_{t-1}| \geq 1, 1 \not\in \mathcal{C}_t \right\} (\alpha - w_0)\gamma_{t-r_1} + \sum_{i \in \mathcal{R}_{t-1} \setminus [1], i \not\in \mathcal{C}_t} \gamma_{t-i} \right).$$

Here $w_0 \in [0, \alpha]$ is an algorithm parameter — we set $w_0 = 0.9$ in all our simulations. r_1 is the index of the first discovery made by LORD*. $(\mathcal{C}_t)$ are a sequence of "conflict sets" that dictate hypothesis indices that the current hypothesis has dependence or "conflict" with. In our local dependence setting, $\mathcal{C}_t = \{t - L, \ldots, t - 1\}$.

B.2 Local dependence simulation details

Each X_t^i is a sample from Beta(a, b) distribution, where we let $a+b = 10^{-2}$, that is shifted and rescaled to be supported on $[-4, 4]$. The following Hoeffding-based process $(M_t^i)_i$ was shown by Waudby-Smith and Ramdas (2023) to be an e-process for random variables bounded in $[\ell, u]$ if $\mathbb{E}[X_t^i] = 0$ for $i \in [N]$.

$$M_t^i = \exp\left(\sum_{j=1}^{i} \lambda_t^j X_t^j - \frac{(\lambda_t^j(u-\ell))^2}{8}\right),$$

for any sequence of $(\lambda_t^j)_{j\in[N]}$ that is predictable, i.e., λ_t^j can be determined by $X_t^1, \ldots, X_t^{j-1}$. We let $\lambda_t^j = \sqrt{8\log(1/(\alpha\gamma_t))/((u-\ell)^2 N)}$ as per Waudby-Smith and Ramdas (2023, eq. 3.6).

Our e-values, and p-values are defined as follows:

$$E_t = M_t^{\tau_t} \text{ and } P_t = \frac{1}{\max_{i\leq N}\ M_t^i},$$

The stopping time τ_t^E defined the in the following recursive fashion:

$$\tau_t = \min\{i \in [N] : M_t^i \geq 1/\widehat{\alpha}_t^{\text{e-LOND}}(i)\} \cup \{N\},$$

where we define $\widehat{\alpha}_t^{\text{e-LOND}}(i)$ to be the test level output by e-LOND after being applied to $(M_1^{\tau_1\wedge i}, \ldots, M_{t-1}^{\tau_{t-1}\wedge i})$, where $\wedge$ denotes minimum. Note that $(M_1^{\tau_1\wedge i}, \ldots, M_{t-1}^{\tau_{t-1}\wedge i})$ can be computed using only the first i samples of the data for the first $t-1$ hypotheses, i.e., $\{X_k^j\}_{j\in[i],k\in[t-1]}$. Hence, these are valid stopping times.

B.3 Sampling WoR simulation details

Let $[\ell, u]$ be the support of the population, and in our case, we set $\ell = -4, u = 4$. Let $P(\mu)$ be the distribution $X = (u-\ell)Y + u$, where $Y \sim \text{Beta}((\mu-\ell)\cdot s/(u-\ell), (u-\mu)\cdot s/(u-\ell))$, i.e., $P(\mu)$ is the Beta distribution scaled to be supported on $[\ell, u]$ with mean μ, and variance scaling factor s (where a smaller s results in population values concentrating at the support limits). Next, take a discrete grid of size $N \times T$ that is uniformly spread over $[0, 1]$, and compute the quantiles of the grid values of $P(\mu)$. We then shift all quantile values below (or above) μ by the same amount, so the mean of the grid quantiles is equal to μ.

The e-values and p-values we use in this setup are derived from the following e-process from Waudby-Smith and Ramdas (2020) for sampling WoR:

$$M_t^i = \exp\left(\sum_{j=1}^{i} \lambda_t^j X_t^j + \mu_t^{j-1}(0) - \frac{(\lambda_t^j(u-\ell))^2}{8}\right),$$

for any predictable sequence $(\lambda_t^i)_{i\in[N]}$ where $\mu_t^i(0) = \frac{1}{N-i+1}\sum_{j=1}^{i} X_t^j$ is an adjustment term for sampling WoR. We also set $\lambda_t^j = \sqrt{8\log(1/(\alpha\gamma_t))/((u-\ell)^2 N)}$ here. We define our e-values and p-values likewise:

$$E_t = M_t^{\tau_t} \text{ and } P_t = \frac{1}{\max_{i\leq N}\ M_t^i},$$

where $\tau_t = \min\{i \in [N] : M_t^i \geq 1/(\alpha\gamma_t)\} \cup \{N\}$ is the first time the $(M_t^i)_{i\in[N]}$ crosses the threshold $1/(\alpha\gamma_t)$ or reaches the maximum sample size N.

C FDR control of e-LOND is sharp

Here we show that there exists a sequence of e-values $(E_1, \ldots, E_t)$ such that the FDR control of e-LOND is sharp.

Theorem 6. *If the discount sequence (γ_t) satisfies $\sum_{t \in \mathbb{N}} \gamma_t = 1$, there exists a joint distribution over a sequence of e-values $(E_t)_{t \in \mathbb{N}}$ such that for every $\varepsilon > 0$, there exists $t' \in \mathbb{N}$ such that* $\mathrm{FDR}(\mathcal{R}_t^{\text{e-LOND}}) > \alpha - \varepsilon$ *for all $t \geq t'$.*

Proof. We write $\mathcal{R}_t$ as shorthand for $\mathcal{R}_t^{\text{e-LOND}}$. We let null be true at every hypothesis, i.e., $\mathcal{H}_0 = \mathbb{N}$, and construct the joint distribution over e-values is characterized as follows:

$$\xi_t \coloneqq \{E_t = (\alpha\gamma_t)^{-1} \text{ and } E_i = 0 \text{ for all } i \neq t\}, \qquad \xi_0 \coloneqq \{E_t = 0 \text{ for all } t \in \mathbb{N}\}$$
$$\mathbb{P}\left(\xi_t\right) = \alpha\gamma_t \text{ for each } t \in \mathbb{N}, \qquad \mathbb{P}\left(\xi_0\right) = 1 - \alpha.$$

Note that ξ_t are disjoint events for $t \in \mathbb{N} \cup \{0\}$, and $\mathbb{P}\left(\xi_0\right) + \sum_{t \in \mathbb{N}} \mathbb{P}\left(\xi_t\right) = 1 - \alpha\alpha \sum_{t \in \mathbb{N}} \gamma_t$. — hence this characterizes a complete distribution over $(E_t)_{t \in \mathbb{N}}$. Further, $\mathbb{E}[E_t] = (\alpha\gamma_t)^{-1} \cdot \mathbb{P}\left(\xi_t\right) = 1$, for each $t \in \mathbb{N}$, so $(E_t)_{t \in \mathbb{N}}$ is provably a sequence of e-values.

We note that $\mathrm{FDP}(\mathcal{R}_t) = \max_{t_1 \in [t]} \mathbf{1}\left\{\xi_{t_1}\right\}$, i.e., the FDP is 1 iff ξ_{t_1} for some $t_1 \in [t]$ occurs. Hence,

$$\mathrm{FDR}(\mathcal{R}_t) = \mathbb{E}\left[\max_{t_1 \in [t]} \mathbf{1}\left\{\xi_{t_1}\right\}\right] = \mathbb{P}\left(\bigcup_{t_1 \in [t]} \xi_{t_1}\right) = \sum_{t_1 \in [t]} \mathbb{P}\left(\xi_{t_1}\right) = \alpha \sum_{t_1 \in [t]} \gamma_{t_1}. \tag{22}$$

Hence, for a fixed $\varepsilon > 0$, if we define $t'(\varepsilon)$ to be the smallest $t \in \mathbb{N}$ such that $\sum_{t_1 \in [t]} \gamma_{t_1} > 1 - (\varepsilon/\alpha)$ — note such a t always exists because (γ_t) is nonnegative and $\sum_{t \in \mathbb{N}} \gamma_t = 1$. We can see as a result of (22), $\mathrm{FDR}(\mathcal{R}_t) > \alpha - \epsilon$ for all $t \geq t'(\epsilon)$. Thus, we have shown our desired result. □

A similar argument can be made to argue that Ue-LOND is sharp as well, as well as FCR control of e-LOND-CI and Ue-LOND-CI.