arXiv:2407.15733v4 [stat.ME] 15 Dec 2025

Admissible online closed testing must employ e-values

Lasse Fischer and Aaditya Ramdas
University of Bremen and Carnegie Mellon University
fischer1@uni-bremen.de, aramdas@cmu.edu

December 16, 2025

Abstract

In contemporary research, data scientists often test an infinite sequence of hypotheses $H_1, H_2, \dots$ one by one, and are required to make real-time decisions without knowing the future hypotheses or data. In this paper, we consider such an online multiple testing problem with the goal of providing simultaneous lower bounds for the number of true discoveries in data-adaptively chosen rejection sets. Employing the recent online closure principle, we show that for this task it is necessary to use an anytime-valid test for each intersection hypothesis. This connects two distinct branches of the literature: online testing of multiple hypotheses (where the hypotheses appear online), and sequential anytime-valid testing of a single hypothesis (where the data for a fixed hypothesis appears online). Motivated by this result, we construct a new online closed testing procedure and a corresponding short-cut with a true discovery guarantee based on multiplying sequential e-values. This general but simple procedure gives uniform improvements over the state-of-the-art methods but also allows to construct entirely new and powerful procedures.

1 Introduction

In the online multiple testing framework a potentially infinite stream of hypotheses $H_1, H_2, \dots$ is tested one by one over time [9, 19]. At each time $t \in \mathbb{N}$, one must decide on the hypothesis H_t without knowing the future hypotheses or data. This setting occurs in the tech industry [22, 27], machine learning [7, 40], open data repositories as used in genomics [2, 26, 1] and other data science tasks where flexible and real-time decision making is required.

A common error metric for a chosen rejection set R_t at time t is the *false discovery proportion*

$$\mathrm{FDP}(R_t) = \frac{\text{Number of true hypotheses in } R_t}{\text{Size of } R_t}. \tag{1}$$

Classical offline and online methods typically control its expectation, the false discovery rate (FDR) [4], below some level $\alpha \in (0, 1)$. However, since the FDP may have high variance, controlling its expectation may not be enough.

In a seminal work, Goeman and Solari [14] proposed to control the tail probabilities of $\mathrm{FDP}(S)$ *simultaneously over all possible sets* S instead. More precisely, they suggest to provide an upper bound $\boldsymbol{q}(S)$ for $\mathrm{FDP}(S)$ such that $\boldsymbol{q}(S) \geq \mathrm{FDP}(S)$ for all S with probability at least $1 - \alpha$.

A major advantage of such simultaneous bounds on the FDP, compared to FDR control, is that the final rejection set(s) can be chosen post-hoc, meaning after looking at the data and calculating the bounds $\boldsymbol{q}(S)$, without violating this error control. In other words, a scientist is allowed to query many (or all) such sets S, examine the reported bounds, and later choose one or a few final sets (and bounds) to report or follow up on. Since many more sets S will be queried than will be rejected, we call these sets S as *query* sets (rather than, say, *rejection* sets).

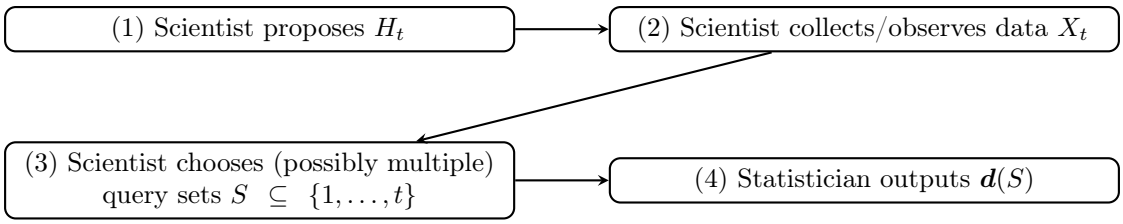


Figure 1: Illustration of using an online procedure with simultaneous true discovery guarantee. At each time t, (1) the scientist proposes a hypothesis H_t for testing, possibly based on the data used for testing the previous hypotheses; (2) the scientist collects the data X_t required for testing H_t which may consist of new data and/or the reuse of old data; (3) the scientist chooses based on the data observed so far (possibly multiple) query sets $S \subseteq \{1, \ldots, t\}$ of interest — if the scientist is interested in rejections, these query sets could also be interpreted as candidates for rejection; (4) the statistician employs the online procedure $\boldsymbol{d}$ to provide lower bounds $\boldsymbol{d}(S)$ for the number of true discoveries in the query sets S that hold simultaneously with probability of at least $1 - \alpha$.

Bounds that hold with high probability are advantageous when an inflated FDP has severe consequences. However, there are also connections between FDR and simultaneous FDP controlling procedures [15, 20, 25]. Following Goeman et al. [16], we consider simultaneous *lower bounds on the number of true discoveries* $\boldsymbol{d}(S) \coloneqq (1 - \boldsymbol{q}(S))|S|$, which is equivalent to bounding the FDP from above but often more convenient. Their work focused on offline multiple testing, while we study this error metric in the online setting.

At each time t, an *online true discovery procedure* $\boldsymbol{d}$ allows the scientist to pick any (and possibly multiple) query sets $S \subseteq \{1, \ldots, t\}$ of interest. Then $\boldsymbol{d}(S)$ instantly provides a lower bound for the number of false hypotheses in S, which holds true with probability at least $1 - \alpha$ simultaneously over all sets that have been queried or that might be queried in the future. Note that this simultaneity permits to stop (or continue) the testing process data-adaptively at any time, e.g. after the 50th discovery. This is an additional benefit compared to online FDR control, which is usually only provided at fixed times [41, 39]. Figure 1 illustrates this process.

A recently popularized approach for online testing of a single hypothesis, where not the hypotheses but the data itself comes in sequentially, is the e-value. An e-value E_t for a hypothesis H_t is a nonnegative random variable which has expected value less than or equal to one if H_t is true. The e-value (or its sequential extension, the e-process) is an alternative to the well-known p-value but more suitable for settings where early stopping of the sampling process or optional continuation is desired [31, 17, 29]. In this paper, we exploit this sequential suitability of the e-value to design online multiple testing procedures with simultaneous true discovery guarantees. This connects two mostly separate areas of the literature: online testing of multiple hypotheses and sequential testing of a single hypothesis. In Supplementary Material S.1 we provide a more detailed description of these existing branches of literature.

1.1 Our contribution

Our contribution can be divided into a theoretical and a methodological part.

Our theoretical contribution. We provide new insights about the central role of e-values in online multiple testing. In particular, we show that in order to derive admissible online procedures with simultaneous true discovery guarantee, we must necessarily employ *anytime-valid tests* for each intersection hypothesis, which in turn *must* employ test martingales [28]. By transitivity, the construction of these online true discovery procedures must rely on test martingales, which are sequential generalizations of e-values. Thus, e-values enter naturally into the construction of

admissible online procedures. Even if an online procedure is constructed solely based on p-values, it must implicitly employ anytime-valid tests and can be reconstructed or improved using e-values.

Our methodological contribution. Guided by our theoretical results, we construct `SeqE-Guard`, a powerful and computationally efficient online true discovery procedure based on multiplying sequential e-values. This improves the existing methodology on several fronts.

1. Due to the flexibility provided by e-values, `SeqE-Guard` is a very general procedure. We show that by plugging specific e-values into `SeqE-Guard`, the online methods of Katsevich and Ramdas [20] and Meah et al. [25], which are all online true discovery procedures we know of, can be uniformly improved. In addition, `SeqE-guard` improves some offline procedures (Section 3.2).

2. The `SeqE-Guard` algorithm operates according to simple rules and is therefore easy to analyze. For instance, `SeqE-Guard` reduces the complex existing methods by Katsevich and Ramdas [20] and Meah et al. [25] with various parameters to simple e-values that can only take two or three different values (Table 1). This facilitates the interpretation and comparison of these methods.

3. The existing online procedures only provide bounds for one query path, meaning a specific nondecreasing sequence of query sets. `SeqE-Guard` allows to analyze all possible query sets simultaneously, without sacrificing power (Corollary 3.2).

4. `SeqE-Guard` does not only improve existing procedures, but can also be used to construct entirely new methods, depending on the choice of the e-values. We investigate the use of `SeqE-Guard` with growth rate optimal (GRO) e-values [30, 17] and propose a hedging approach to increase the power to detect false hypotheses (Section 3.3).

1.2 Paper outline

In Section 2, we define the online setting formally and recap concepts like coherence (Section 2.1) and (online) closed testing (Section 2.2). Afterwards, we introduce a general approach to online true discovery guarantee based on test martingales and prove that every procedure must be constructed in that way (Section 2.3). In Section 3, we consider online true discovery guarantee with sequential e-values and propose our `SeqE-Guard` algorithm for this task. By plugging specific sequential e-values into `SeqE-Guard` we immediately obtain uniform improvements of the state-of-the-art methods by Katsevich and Ramdas [20], Meah et al. [25] (Section 3.2). Separately, we consider `SeqE-Guard` with GRO e-values in Section 3.3. In Section 4, we perform simulations to compare the proposed methods and to quantify the gain in power obtained by our improvements[1].

2 Online true discovery guarantee

In this section, we introduce notation and recall concepts like simultaneous (online) true discovery guarantee and (online) closed testing. Then, we prove that any admissible procedure for delivering an online true discovery guarantee must rely on test martingales (sequential generalizations of e-values).

We consider the general online multiple testing setting described in [8]. Let $(\Omega, \mathbb{F})$, where $\mathbb{F} = (\mathcal{F}_i)_{i\in\mathbb{N}_0}$, be a filtered measurable space and $\mathcal{P}$ some set of probability distributions on $(\Omega, \mathbb{F})$. The σ-field $\mathcal{F}_i$ defines the information that can be used for testing hypothesis H_i ($\mathcal{F}_0 = \{\emptyset, \Omega\}$). Hence, in online multiple testing every hypothesis test is only allowed to use some partial information which is increasing over time. One can think of $\mathcal{F}_i$ as the data that is available at time i. However, we may also include external randomization or coarsen the filtration. For example, many existing works

[1]The code for the simulations is available at `github.com/fischer23/online_true_discovery`.

on online multiple testing consider $\mathcal{F}_i = \sigma(P_1, \dots, P_i)$ [9, 19], where each P_j is a p-value calculated for hypothesis H_j. We would also like to point out that the hypotheses do not have to be prespecified. Hence, in practice one is allowed to data-adaptively construct each H_i based on $\mathcal{F}_{i-1}$.

We assume that the data follows some unknown distribution $\mathbb{P} \in \mathcal{P}$. A null hypothesis $H \subseteq \mathcal{P}$ is a collection of probability distributions; we are effectively testing whether $\mathbb{P} \in H$. When $\mathbb{P} \in H$, we say that H is true null, and otherwise we call it a false null. We define $I_0^{\mathbb{P}} := \{i \in \mathbb{N} : \mathbb{P} \in H_i\}$ and $I_1^{\mathbb{P}} := \mathbb{N} \setminus I_0^{\mathbb{P}}$ as the index sets of true and false null hypotheses, respectively.

As defined in [16, 8], a procedure with *simultaneous true discovery guarantee* is a random function $\boldsymbol{d} : 2^{\mathbb{N}_f} \to \mathbb{N} \cup \{0\}$, where $2^{\mathbb{N}_f}$ is the set of all finite subsets of $\mathbb{N}$ (analogously we use $2^{\mathbb{N}-f}$ for the set of all infinite subsets of $\mathbb{N}$), such that for all $\mathbb{P} \in \mathcal{P}$:

$$\mathbb{P}(\boldsymbol{d}(S) \leq |S \cap I_1^{\mathbb{P}}| \text{ for all } S \in 2^{\mathbb{N}_f}) \geq 1 - \alpha. \tag{2}$$

Clearly, if $\boldsymbol{d}$ always outputs 0, it achieves simultaneous true discovery guarantee. Thus, implicitly, the larger $\boldsymbol{d}$ is, the better (here, larger is meant componentwise; $\boldsymbol{d} \geq \boldsymbol{d}'$ if $\boldsymbol{d}(S) \geq \boldsymbol{d}'(S)$ for all $S \in 2^{\mathbb{N}_f}$). It should be noted that the sets S in (2) can be considered as fixed because of the quantification "for all S" (in contrast to classical FDR control), but the procedure $\boldsymbol{d}$ depends on the data and thus the probability is taken with regard to $\boldsymbol{d}$. However, the simultaneous guarantee of $\boldsymbol{d}$ particularly allows to choose query sets S arbitrarily based on the data while ensuring that $\boldsymbol{d}(S) \leq |S \cap I_1^{\mathbb{P}}|$ with probability at least $1 - \alpha$.

$\boldsymbol{d}$ is called an *online* true discovery procedure if $\boldsymbol{d}(S)$ is measurable with respect to $\mathcal{F}_{\max(S)}$ for all $S \in 2^{\mathbb{N}_f}$ [8]. This ensures that at any time $t \in \mathbb{N}$ the procedure $\boldsymbol{d}$ provides a lower bound for the number of false hypotheses (or, equivalently, true discoveries) in every set $S \subseteq \{1, \dots, t\}$ with $\max(S) = t$ that remains the same no matter how many hypotheses will be tested in the future. See also Figure 1 for an illustration of using online true discovery procedures.

Note that one does not have to consider an infinite number of hypotheses but could just stop at some finite time $i \in \mathbb{N}$ by setting $H_j = \mathcal{P}$ for all $j > i$ and $\boldsymbol{d}(S) = \boldsymbol{d}(S \cap \{1, \dots, i\})$ for all $S \in 2^{\mathbb{N}_f}$ with $\max(S) > i$. With this, the online setting becomes classical offline testing in the case of $\mathcal{F}_1 = \mathcal{F}_2 = \dots$ and thus online multiple testing can be seen as a true generalization of classical multiple testing [8]. Although we are mainly interested in the strict online case $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$, this implies that all online procedures in this paper also apply in the offline setting.

Further note that simultaneous true discovery procedures generalize many existing error rates, such as k-FWER or false discovery exceedance (FDX) [16]. In particular, this shows that requiring simultaneous control *over all* sets $S \in 2^{\mathbb{N}_f}$ is not conservative. Intuitively, this is because a procedure $\boldsymbol{d}$ that assigns lower bounds only to a subset G of $2^{\mathbb{N}_f}$ can be interpolated to a procedure on the entire set $2^{\mathbb{N}_f}$ (see Goeman et al. [15]).

In Figure 2, we connect the results that will be given in the rest of the section and the related works [28, 8]. The left-hand side provides a general approach to construct coherent online true discovery procedures based on test martingales and the right-hand side proves that every admissible online procedure must be constructed in that way. While most of the upper and lower connections are known, the main insights provided by this paper are the middle connections between increasing families of online intersection tests and anytime-valid tests, connecting two distinct branches of the literature. All terms and results will be clarified in the following subsections from top to bottom, starting with *coherent online procedures*.

2.1 Coherent online true discovery procedures

An important property of multiple testing procedures is *coherence* [11, 32]. A true discovery procedure $\boldsymbol{d}$ is called *coherent* [16], if for all disjoint $S, U \in 2^{\mathbb{N}_f}$,

$$\boldsymbol{d}(S) + \boldsymbol{d}(U) \leq \boldsymbol{d}(S \cup U) \leq \boldsymbol{d}(S) + |U|. \tag{3}$$

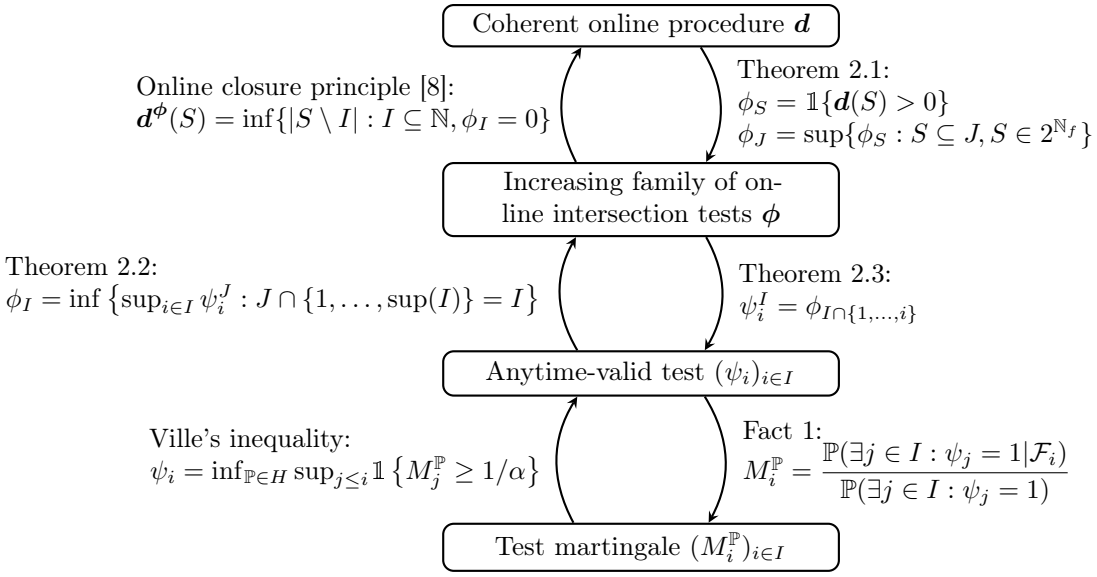


Figure 2: Illustration of the relation between online true discovery procedures, increasing families of online intersection tests, anytime-valid tests and test martingales. The left path from bottom to top provides a general approach for the construction of coherent online procedures with true discovery guarantee based on test martingales. The right path from top to bottom shows that coherent online procedure implicitly define test martingales. Taken together, the loop shows, in principle, how to improve any given coherent online procedure: we take the right path down, and then take the left path up. If the original procedure is admissible, following the loop leaves the procedure unchanged.

Coherence ensures consistent decisions or bounds of the multiple testing procedure and is therefore a desirable property. A procedure $\boldsymbol{d}$ is admissible if there is no other procedure $\tilde{\boldsymbol{d}}$ that uniformly improves $\boldsymbol{d}$, where $\tilde{\boldsymbol{d}}$ is said to uniformly improve $\boldsymbol{d}$, if $\tilde{\boldsymbol{d}} \ge \boldsymbol{d}$ and $\mathbb{P}(\tilde{\boldsymbol{d}}(S) > \boldsymbol{d}(S)) > 0$ for at least one $\mathbb{P} \in \mathcal{P}$ and at least one $S \in 2^{\mathbb{N}_f}$. Equivalently, $\boldsymbol{d}$ is admissible if $\tilde{\boldsymbol{d}} \ge \boldsymbol{d}$ implies $\tilde{\boldsymbol{d}} = \boldsymbol{d}$.

Goeman et al. [16] showed that in the offline setting, all admissible true discovery procedures must be coherent. However, it turns out that this result is not true in the online case.

Example 1. *Consider a setting with only two hypotheses H_1 and H_2 with independent p-values P_1 and P_2 that are uniformly distributed under the null hypothesis. Let $\boldsymbol{d}(\{1\}) = 1$, if $P_1 \le \alpha/2$ (and 0 otherwise), $\boldsymbol{d}(\{2\}) = 1$, if $P_2 \le \alpha/2$ (and 0 otherwise) and*

$$\boldsymbol{d}(\{1,2\}) = \begin{cases} 2, & (P_1 \le \alpha/2 \wedge P_2 \le \alpha) \vee (P_2 \le \alpha/2 \wedge P_1 \le \alpha) \\ 1, & (P_1 \le \alpha/2 \wedge P_2 > \alpha) \vee (P_2 \le \alpha/2 \wedge P_1 > \alpha) \\ 0, & \textit{otherwise.} \end{cases}$$

Then $\boldsymbol{d}$ is an incoherent online procedure, since $\boldsymbol{d}(\{1,2\}) = 2$ and $\boldsymbol{d}(\{1\}) = 0$ if $P_2 \le \alpha/2$ and $\alpha/2 < P_1 \le \alpha$. However, it is not possible to improve $\boldsymbol{d}$ by a coherent online procedure $\tilde{\boldsymbol{d}} \ge \boldsymbol{d}$. To see this, note that in order to be coherent, $\tilde{\boldsymbol{d}}(\{1\})$ needs to equal 1, if $P_2 \le \alpha/2 \wedge P_1 \le \alpha$. Since $\tilde{\boldsymbol{d}}$ is an online procedure, $\tilde{\boldsymbol{d}}(\{1\})$ must not use information about P_2, so we must have $\tilde{\boldsymbol{d}}(\{1\}) = 1$, if $P_1 \le \alpha$. However, this implies that for all $\mathbb{P} \in H_1 \cap H_2$ such that true discovery guarantee is not provided, $\mathbb{P}(\{\tilde{\boldsymbol{d}}(\{1\}) = 0\} \cap \{\tilde{\boldsymbol{d}}(\{2\}) = 0\}) \le \mathbb{P}(\{P_1 > \alpha\} \cap \{P_2 > \alpha/2\}) = (1-\alpha)(1-\alpha/2) < 1-\alpha$.

Still, it is sensible to focus on coherent procedures as incoherent results are difficult to interpret.

2.2 Online closed testing

The closure principle was originally proposed for FWER control [24]. Genovese and Wasserman [12, 13] proposed a similar approach for simultaneous true discovery guarantee, which was later formulated in terms of closed testing by Goeman and Solari [14], Goeman et al. [15]. For each intersection hypothesis $H_I = \bigcap_{i \in I} H_i$, let ϕ_I be an intersection test and $\boldsymbol{\phi} = (\phi_I)_{I \subseteq \mathbb{N}}$ denote the family of intersection tests. Henceforth, it is understood that all tests are α-level tests, meaning $\mathbb{P}(\phi_I = 1) \leq \alpha$ for all $\mathbb{P} \in H_I$. Genovese and Wasserman [12, 13] showed that

$$\boldsymbol{d}^{\boldsymbol{\phi}}(S) \coloneqq \inf\{|S \setminus I| : I \subseteq \mathbb{N}, \phi_I = 0\} \qquad (S \in 2^{\mathbb{N}_f}) \tag{4}$$

provides simultaneous true discovery guarantee over all $S \in 2^{\mathbb{N}_f}$. This follows from the fact that $\boldsymbol{d}^{\boldsymbol{\phi}}(S) > |S \cap I_1^{\mathbb{P}}|$ for any $S \in 2^{\mathbb{N}_f}$ implies $\phi_{I_0^{\mathbb{P}}} = 1$, which happens with probability at most α.

Since we adopt the closed testing framework by Goeman and Solari [14], Goeman et al. [15], we also refer to $\boldsymbol{d}^{\boldsymbol{\phi}}$ as closed procedure. Technically, the aforementioned works only considered finitely many hypotheses, but the method and its guarantees extend to a countable number of hypotheses [8], and so we present that version for easier connection to the online setting.

Goeman et al. [16] proved an important result: every (offline) coherent true discovery procedure is equivalent to or uniformly improved by a closed procedure of the form (4). Therefore, the closure principle allows to construct and analyze all admissible true discovery procedures based on single tests for the intersection hypotheses which are usually much easier to handle. Fischer et al. [8] showed that the closed procedure $\boldsymbol{d}^{\boldsymbol{\phi}}$ is an online procedure, if the following assumptions are fulfilled:

1. Every ϕ_I, $I \subseteq \mathbb{N}$ is an *online intersection test*, meaning ϕ_I is measurable with respect to $\mathcal{F}_{\sup(I)}$.

2. The family of intersection tests $\boldsymbol{\phi} = (\phi_I)_{I \subseteq \mathbb{N}}$ is *increasing*[2]: for all $i \in \mathbb{N}$ and $I \subseteq \{1, \dots, i\}$,

$$\phi_I \leq \phi_{I \cup K} \text{ for all } K \subseteq \{k \in \mathbb{N} : k > i\}. \tag{5}$$

This follows from the fact that the closed procedure $\boldsymbol{d}^{\boldsymbol{\phi}}$ for increasing $\boldsymbol{\phi}$ can be written as

$$\boldsymbol{d}^{\boldsymbol{\phi}}(S) \coloneqq \inf\{|S \setminus I| : I \subseteq \{1, \dots, \max(S)\}, \phi_I = 0\} \qquad (S \in 2^{\mathbb{N}_f}). \tag{6}$$

A closed procedure $\boldsymbol{d}^{\boldsymbol{\phi}}$, where $\boldsymbol{\phi}$ satisfies the conditions 1 and 2, is called an *online closed procedure*. Every online closed procedure is a coherent online procedure, which follows immediately from the same result in the offline case [16]. Fischer et al. [8] proved that all online procedures with FWER control can be written as a closed procedure where the intersection tests satisfy these two conditions. Hence we know that the closure principle is admissible for offline true discovery control [16] and online FWER control [8], but not yet for coherent online true discovery control. We close this gap by showing that any coherent online procedure can be recovered or improved by an online closed procedure.

Theorem 2.1. *Let $\boldsymbol{d}$ be a coherent online true discovery procedure. Define*

$$\phi_S = \mathbb{1}\{\boldsymbol{d}(S) > 0\} \ \forall S \in 2^{\mathbb{N}_f} \quad \text{and} \quad \phi_J = \sup\{\phi_S : S \subseteq J, S \in 2^{\mathbb{N}_f}\} \ \forall J \in 2^{\mathbb{N}_{-f}}. \tag{7}$$

Then $\boldsymbol{\phi} = (\phi_I)_{I \subseteq \mathbb{N}}$ is an increasing family of online intersection tests and $\boldsymbol{d}^{\boldsymbol{\phi}} \geq \boldsymbol{d}$.

[2]Fischer et al. [8] used the term *predictable* for (5), but we use *increasing* in order to avoid confusion with the measure-theoretic definition of predictability.

Proof. The simultaneous true discovery guarantee of $\boldsymbol{d}$ implies that the ϕ_I define intersection tests. Furthermore, since $\boldsymbol{d}$ is coherent, it follows that $\phi_S \leq \phi_I$ if $S \subseteq I$, which particularly implies that $\boldsymbol{\phi}$ is increasing. In addition, ϕ_S is an online intersection test since $\boldsymbol{d}$ is an online procedure. Therefore, it only remains to show that $\boldsymbol{d}(S) \leq \boldsymbol{d}^{\boldsymbol{\phi}}(S)$ for all $S \in 2^{\mathbb{N}_f}$. Suppose $\boldsymbol{d}(S) = s$. Due to the coherence of $\boldsymbol{d}$, we have $\phi_J = 1$ for all $J \subseteq S$ with $|J| > |S| - s$. The coherence further implies that $\phi_I = 1$ for all $I \subseteq \mathbb{N}$ with $|S \cap I| > |S| - s$ and hence $\boldsymbol{d}^{\boldsymbol{\phi}}(S) \geq s$. □

With this result, we can focus on online closed procedures when considering coherent online true discovery guarantees. In the following subsection, we introduce our main result and show that to construct online closed procedures, it is necessary to define anytime-valid tests and therefore one must rely on test martingales. This yields the connected graph in Figure 2 and motivates the construction of e-value based online closed procedures, which we will focus on afterwards.

Note that Theorem 2.1 particularly holds in the offline case $\mathcal{F}_1 = \mathcal{F}_2 = \dots$ and therefore immediately yields the aforementioned result by Goeman et al. [16] as a corollary.

Remark 1. *There are cases where the closed procedure based on the intersection tests defined in* (7) *dominates the original procedure. For example, consider three hypotheses and the coherent procedure* $\boldsymbol{d}$ *with* $\boldsymbol{d}(\{1,2,3\}) = 1$, $\boldsymbol{d}(\{1,2\}) = 1$, $\boldsymbol{d}(\{1,3\}) = 1$, $\boldsymbol{d}(\{2,3\}) = 1$ *and* $\boldsymbol{d}(\{i\}) = 0$ *for* $i \in \{1,2,3\}$. *The corresponding closed procedure would give the same bounds except for further concluding that* $\boldsymbol{d}(\{1,2,3\}) = 2$. *This makes sense, since if there is at least one true discovery in* $\{1,2\}$, *one in* $\{1,3\}$ *and one in* $\{2,3\}$, *there should be at least two true discoveries in* $\{1,2,3\}$.

2.3 Admissible coherent online procedures for true discovery guarantees must rely on test martingales

Anytime-valid tests are used to sequentially test a single null hypothesis with increasing data. A famous example is the sequential probability ratio test (SPRT) by Wald [36] for simple hypotheses, which rejects the null hypothesis as soon as the likelihood ratio of the alternative distribution against the null distribution is above some predefined threshold. In recent years, anytime-valid tests have received renewed attention with a particular focus on general composite hypotheses [29]. For example, Ramdas et al. [28] have shown that every anytime-valid test has a similar form to the aforementioned SPRT, in which the likelihood ratio is replaced by the more general concept of e-processes, which in turn are based on test martingales. In this section, we make connections between increasing families of online intersection tests and anytime-valid tests, which then reveal close relations of online procedures with true discovery guarantee and test martingales.

Suppose we have an anytime-valid test $(\psi_i^I)_{i \in I}$ for each intersection hypothesis H_I. Following [29], $(\psi_i^I)_{i \in I}$ is an anytime-valid test for H_I, if ψ_i^I is measurable with respect to $\mathcal{F}_i$ and we have $\mathbb{P}(\exists i \in I : \psi_i^I = 1) \leq \alpha$ for all $\mathbb{P} \in H_I$. The following theorem shows how we can use such anytime-valid tests to construct an increasing family of online intersection tests.

Theorem 2.2. *For all* $I \subseteq \mathbb{N}$ *let* $(\psi_i^I)_{i \in I}$ *be an anytime-valid test for* H_I. *Then* $\boldsymbol{\phi} = (\phi_I)_{I \subseteq \mathbb{N}}$, *where*

$$\phi_I = \inf \left\{ \sup_{i \in I} \psi_i^J : J \cap \{1, \dots, \sup(I)\} = I \right\}, \tag{8}$$

is an increasing family of online intersection tests. In particular, for infinite $I \subseteq \mathbb{N}$, $\phi_I = \sup_{i \in I} \psi_i^I$.

Proof. Since $(\psi_i^I)_{i \in I}$, $I \subseteq \mathbb{N}$, is an anytime-valid test for H_I, it immediately follows that ϕ_I is an online intersection test. Furthermore, for $i \in \mathbb{N}$, $I \subseteq \{1, \dots, i\}$ and $K \subseteq \{k \in \mathbb{N} : k > i\}$ it holds that

$J \cap \{1, \dots, \sup(I \cup K)\} = I \cup K$ implies that $J \cap \{1, \dots, \sup(I)\} = I$. Therefore,

$$\begin{aligned}\phi_I &= \inf\left\{\sup_{i \in I} \psi_i^J : J \cap \{1, \dots, \sup(I)\} = I\right\} \\ &\leq \inf\left\{\sup_{i \in I} \psi_i^J : J \cap \{1, \dots, \sup(I \cup K)\} = I \cup K\right\} \\ &\leq \inf\left\{\sup_{i \in I \cup K} \psi_i^J : J \cap \{1, \dots, \sup(I \cup K)\} = I \cup K\right\} = \phi_{I \cup K},\end{aligned}$$

showing that $\boldsymbol{\phi}$ is increasing. □

In the following theorem, we show a converse relationship, meaning that increasing families of online intersection tests implicitly define anytime-valid tests for the intersection hypotheses. In addition, we prove that every increasing family of online intersection tests can be constructed by anytime-valid tests using (8).

Theorem 2.3. *Let $\boldsymbol{\phi} = (\phi_I)_{I \subseteq \mathbb{N}}$ be an increasing family of online intersection tests. Define*

$$\psi_i^I = \phi_{I \cap \{1, \dots, i\}}. \tag{9}$$

Then $(\psi_i^I)_{i \in I}$ is an anytime-valid test for H_I for all $I \subseteq \mathbb{N}$. Furthermore, let $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_I)_{I \subseteq \mathbb{N}}$ be defined by $(\psi_i^I)_{i \in I}$, $I \subseteq \mathbb{N}$, using (8). Then $\tilde{\phi}_S = \phi_S$ for all $S \in 2^{\mathbb{N}_f}$ and $\boldsymbol{d}^{\tilde{\boldsymbol{\phi}}} = \boldsymbol{d}^{\boldsymbol{\phi}}$.

Proof. Since $\phi_{I \cap \{1, \dots, i\}}$ is an online intersection test, we have that ψ_i^I is measurable with respect to $\mathcal{F}_i$. Furthermore, since $\boldsymbol{\phi}$ is increasing, it holds that $\mathbb{P}(\exists i \in I : \psi_i^I = 1) = \mathbb{P}(\phi_I = 1) \leq \alpha$ for all $\mathbb{P} \in H_I$. Hence, $(\psi_i^I)_{i \in I}$ is an anytime-valid test for H_I with respect to the filtration $(\mathcal{F}_i)_{i \in I}$. Now let $\tilde{\boldsymbol{\phi}}$ be defined by (8). Then for all $S \in 2^{\mathbb{N}_f}$,

$$\begin{aligned}\tilde{\phi}_S &= \inf\left\{\sup_{i \in S} \psi_i^J : J \cap \{1, \dots, \sup(S)\} = S\right\} \\ &= \inf\left\{\sup_{i \in S} \phi_{J \cap \{1, \dots, i\}} : J \cap \{1, \dots, \sup(S)\} = S\right\} \\ &= \sup_{i \in S} \phi_{S \cap \{1, \dots, i\}} = \phi_S,\end{aligned}$$

where the last equality follows since $\boldsymbol{\phi}$ is increasing. Note that this implies $\boldsymbol{d}^{\tilde{\boldsymbol{\phi}}} = \boldsymbol{d}^{\boldsymbol{\phi}}$, since $\tilde{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}$ are increasing and therefore $\boldsymbol{d}^{\tilde{\boldsymbol{\phi}}}(S)$ and $\boldsymbol{d}^{\boldsymbol{\phi}}(S)$, $S \in 2^{\mathbb{N}_f}$, can be determined solely based on $\tilde{\phi}_I$ and ϕ_I, respectively, with $I \subseteq S$. □

Together, Theorems 2.1 and 2.3 imply that in order to define admissible coherent online true discovery procedures, we need to construct anytime-valid tests for the intersection hypotheses. In addition, they imply that every coherent online true discovery procedure implicitly defines nontrivial anytime-valid tests by (7) and (9) for each intersection hypothesis. These results are particularly interesting, since Ramdas et al. [28] gave a precise characterization of anytime-valid tests, proving that every anytime-valid test can be reconstructed or uniformly improved using test martingales. To state their result more precisely, we need to introduce some terminology.

An anytime-valid test $(\tilde{\psi}_i)_{i \in I}$ uniformly improves $(\psi_i)_{i \in I}$, if $\tilde{\psi}_i \geq \psi_i$ for all $i \in I$ and $\mathbb{P}(\tilde{\psi}_i > \psi_i) > 0$ for some $i \in I$ and $\mathbb{P} \in \mathcal{P}$. Furthermore, a nonnegative process $(M_i)_{i \in I \cup \{0\}}$ adapted to the filtration $(\mathcal{F}_i)_{i \in I \cup \{0\}}$, where $\mathcal{F}_0 = \emptyset$, is a test (super)martingale for $\mathbb{P}$, if $M_0 \overset{(\leq)}{=} 1$ and $\mathbb{E}_{\mathbb{P}}[M_i | \mathcal{F}_{i-}] \overset{(\leq)}{=} M_{i-}$ for all $i \in I$, where $i- = \max\{j \in I \cup \{0\} : j < i\}$. We call $(M_i)_{i \in I \cup \{0\}}$ a test (super)martingale for a null hypothesis H_I, if the above holds for all $\mathbb{P} \in H_I$. In the following we formally state the result by Ramdas et al. [28] (adapted to our setup) and provide a self-contained proof.

Fact 1. *Let $(\psi_i)_{i\in I}$ be an anytime-valid test for H_I, $I \subseteq \mathbb{N}$, and define for all $\mathbb{P} \in H_I$:*

$$M_i^{\mathbb{P}} := \frac{\mathbb{P}(\exists j \in I : \psi_j = 1 | \mathcal{F}_i)}{\mathbb{P}(\exists j \in I : \psi_j = 1)} \quad (i \in I \cup \{0\})$$

with $M_i^{\mathbb{P}} = 1$ if the denominator equals zero. Then $(M_i^{\mathbb{P}})_{i\in I\cup\{0\}}$ is a test martingale for $\mathbb{P}$. Furthermore, let

$$\tilde{\psi}_i = \inf_{\mathbb{P}\in H_I} \sup_{j\le i} \mathbb{1}\{M_j^{\mathbb{P}} \ge 1/\alpha\} \quad (i \in I).$$

Then $(\tilde{\psi}_i)_{i\in I}$ is an anytime-valid test for H_I and either equals or uniformly improves $(\psi_i)_{i\in I}$.

Proof. The tower property for conditional expectations implies that $(M_i^{\mathbb{P}})_{i\in I\cup\{0\}}$ is a test martingale for $\mathbb{P}$ and Ville's inequality shows that $(\tilde{\psi}_i)_{i\in I}$ is an anytime-valid test for H_I. Furthermore, $\psi_i = 1$ implies that $\mathbb{P}(\exists j \in I : \psi_j = 1|\mathcal{F}_i) = 1$ and therefore $M_i^{\mathbb{P}} \ge 1/\alpha$ for all $\mathbb{P} \in H_I$. □

Fact 1 closes the loop illustrated in Figure 2. This particularly shows that every coherent online true discovery procedure $\boldsymbol{d}$ implicitly constructs test martingales for the intersection hypotheses. Furthermore, taking the left path in Figure 2 with these test martingales always yields a procedure which either equals or uniformly improves the original procedure $\boldsymbol{d}$.

3 Online true discovery guarantee with sequential e-values

In the previous section we showed that we need to construct anytime-valid tests, and thus test martingales, for the intersection hypotheses when constructing coherent online procedures with a true discovery guarantee. This general martingale-based approach is illustrated in the left path from bottom to top in Figure 2. Since every step involves taking the infimum over a large set, it seems computationally inefficient. However, in practice one can avoid this by using the same test martingale for all $\mathbb{P} \in H$, the same anytime-valid test $(\psi_i^J)_{i\in I}$ for all J with $J \cap \{1, \dots, \sup(I)\} = I$ and constructing $\boldsymbol{\phi}$ in a way that permits a short-cut for $\boldsymbol{d}^{\boldsymbol{\phi}}$. In this section, we apply all this to derive a computationally efficient and powerful online true discovery procedure based on sequential e-values. We first introduce our general `SeqE-Guard` algorithm (Section 3.1), then show that it allows to uniformly improve all existing online true discovery procedures by using specific sequential e-values (Section 3.2) and finally analyze its behavior when using GRO e-values (Section 3.3).

3.1 The `SeqE-Guard` algorithm

Let $(M_t)_{t\in\mathbb{N}_0}$ be a test supermartingale with respect to $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$ for some hypothesis H. We can break down $(M_t)_{t\in\mathbb{N}_0}$ into its individual factors $E_t = \frac{M_t}{M_{t-1}}$, $t > 0$, with the convention $0/0 = 0$. Due to the supermartingale property, E_t is nonnegative, measurable with respect to $\mathcal{F}_t$ and

$$\mathbb{E}_{\mathbb{P}}[E_t|\mathcal{F}_{t-1}] = \mathbb{E}_{\mathbb{P}}[M_t/M_{t-1}|\mathcal{F}_{t-1}] = \mathbb{E}_{\mathbb{P}}[M_t|\mathcal{F}_{t-1}]/M_{t-1} \le 1 \quad (\mathbb{P} \in H). \tag{10}$$

Hence, each of these random variables E_t is an e-value for H conditional on the past, sometimes called a sequential e-value in the literature [34]. Thus, every test supermartingale can be written as the product of sequential e-values. Conversely, every product of sequential e-values defines a test supermartingale (just multiply M_{t-1} on both sides of (10)). For these reasons, sequential e-values are potentially the perfect tool to define online closed procedures, which is why we will focus on them in the following.

Assume that for each hypothesis H_t an e-value E_t is available and the e-values $(E_t)_{t\in\mathbb{N}}$ are sequentially valid with respect to $(\mathcal{F}_t)_{t\in\mathbb{N}_0}$, meaning $E_t \in \mathcal{F}_t$ and $\mathbb{E}_{\mathbb{P}}[E_t|\mathcal{F}_{t-1}] \le 1$ for all $\mathbb{P} \in H_t$. For

each intersection hypothesis H_I, we construct a process $(W_I^t)_{t\in I\cup\{0\}}$ with $W_I^0 = 1$ and

$$W_I^t = \prod_{i\in I\cap\{1,\dots,t\}} E_i \qquad (t \in I). \tag{11}$$

Following the above argumentation, $(W_I^t)_{t\in I\cup\{0\}}$ is a test supermartingale for H_I with respect to $(\mathcal{F}_t)_{t\in I\cup\{0\}}$. By Ville's inequality [18], it follows that

$$\phi_I = \mathbb{1}\{\exists t \in I : W_I^t \geq 1/\alpha\} = \mathbb{1}\left\{\sup_{t\in I} W_I^t \geq 1/\alpha\right\} \tag{12}$$

is an intersection test. Furthermore, ϕ_I is an online intersection test and $\boldsymbol{\phi} = (\phi_I)_{I\subseteq\mathbb{N}}$ is increasing such that the closed procedure $\boldsymbol{d}^{\boldsymbol{\phi}}$ (4) is indeed an online procedure (see Section 2.2). Note that due to the supremum involved in Ville's inequality, the intersection tests ϕ_I, $I \subseteq \mathbb{N}$, are not symmetric, and thus very different from the ones considered by Tian et al. [33].

Sequential e-values arise naturally in a variety of settings [30, 17, 29, 35]. For example, note that many online multiple testing tasks are performed in an adaptive manner. That means, the hypotheses to test, the design of a study or the strategy to calculate an e-value, depend on the data observed so far. This is very natural, since the hypotheses are tested over time and therefore the statistician performing these tests automatically learns about the context of the study and the true distribution of the data during the testing process. Indeed, to avoid such data-adaptive designs one would need to ignore all the previous data when making design decisions for the future testing process and therefore could be required to prespecify all hypotheses that are going to be tested, the sample size for these tests, decide which tests to apply and fix many further design parameters at the very beginning of the testing process. This would take away much of the flexibility of online multiple testing procedures. However, if one uses past information for the design of an e-value and calculates the e-value on the same data that design information is based on, one would need to have knowledge about the conditional distribution of the e-value given that design information. This is often not available. Therefore, one usually uses independent and fresh data for each of the e-values. In this way, no matter in which way an e-value depends on the past data, it remains valid conditional on it. Consequently, we obtain sequential e-values well suited to our online true discovery procedure.

Computational shortcuts. One problem with procedures based on the (online) closure principle is that at each step $t \in \mathbb{N}$ up to 2^{t-1} additional intersection tests must be considered. However, for specific intersection tests this number can be reduced drastically. In Algorithm 1 we introduce a shortcut for the intersection tests in (12) that only requires one calculation per individual hypothesis but provides the same bounds as the entire closed procedure would. We call this Algorithm `SeqE-Guard` (guarantee for true discoveries with sequential e-values).

The `SeqE-Guard` algorithm will provide a sequence of lower bounds d_t on the number of true discoveries in an adaptively chosen sequence of query sets $S_t \subseteq \{1,\dots,t\}$, simultaneously over all t. The S_t and d_t sequences are (nonstrictly) increasing with t since by default the statistician only decides at step t whether to include the index t within S_t or not (based on $E_1,\dots,E_t$), but typically does not omit hypotheses that were deemed interesting at an earlier point (see Remark below). So it is understood below that $S_t \supseteq S_{t-1}$ for all t. As the multiple testing process continues, the statistician can report/announce one or more (S_t, d_t) pairs that they deem interesting thus far, and our theorem below guarantees that all such announcements will be accurate with high probability.

A short description of `SeqE-Guard`. At each step $t \in \mathbb{N}$, the statistician decides based on the e-values $E_1,\dots,E_t$ whether the index t should be included in the query set S_t. If it is to be included, `SeqE-Guard` calculates the product of all e-values in S_t (that were not already excluded) and all e-values smaller than 1 that are not in S_t. If that product is larger than or equal to $1/\alpha$, we increase

the true discovery bound by 1 and *exclude* the current largest e-value from the future analysis. In Supplementary Material S.2 we provide a simple calculation example for the `SeqE-Guard` algorithm.

Theorem 3.1. *Let $E_1, E_2, \dots$ be sequential e-values for $H_1, H_2, \dots$. The true discovery bounds d_t for S_t defined by* ***`SeqE-Guard`*** *(Algorithm 1) are the same as the ones obtained by the online closure principle with the intersection tests in* (12)*. In particular, S_t and d_t satisfy for all $\mathbb{P} \in \mathcal{P}$: $\mathbb{P}(d_t \leq |S_t \cap I_1^{\mathbb{P}}| \text{ for all } t \in \mathbb{N}) \geq 1 - \alpha$.*

The main idea for the proof of Theorem 3.1 is that the bound $\boldsymbol{d}^{\boldsymbol{\phi}}(S_t)$ is given by $|S_t \setminus I|$, where $I \subseteq \{1, \dots, t\}$ is the index set that contains the most possible indices of S_t while guaranteeing that $W_I^s < 1/\alpha$ for all $s \leq t$ (see (6)). The `SeqE-guard` algorithm finds this set I by including all indices $i \in \{1, \dots, t\} \setminus S_t$ with $E_i < 1$ (since those e-values reduce the product) and excluding the index $i \in I \cap S_t$ of the largest e-value in case of $W_I^t \geq 1/\alpha$. Since the full proof is quite long, we have deferred it to Supplementary Material S.7.

Note that closed procedures provide simultaneous true discovery guarantee simultaneously over all $S \in 2^{\mathbb{N}_f}$, while `SeqE-Guard` only gives a simultaneous lower bound on the number of true discoveries for a path of query sets $(S_t)_{t \in \mathbb{N}}$ with $S_1 \subseteq S_2 \subseteq \dots$. However, at some time $t \in \mathbb{N}$ one could also obtain a lower bound for the number of true discoveries in any other set $S \subseteq \{1, \dots, t\}$ that is not on the path. In particular, Theorem 3.1 shows that the bound $\boldsymbol{d}^{\boldsymbol{\phi}}(S)$, with $\boldsymbol{\phi}$ from (12), can be calculated in time linear in $\max(S)$ for any S by setting $S_t = S \cap \{1, \dots, t\}$.

We formulated `SeqE-Guard` for single query paths due to computational convenience and as we think this reflects the proceeding in many applications. This was also done in previous works on online true discovery guarantee [20, 25]. In the following corollary, we capture the guarantee of `SeqE-Guard` for multiple query paths.

Corollary 3.2. *Let $(d_t^{(1)})_{t \in \mathbb{N}}, (d_t^{(2)})_{t \in \mathbb{N}}, \dots$ be obtained by applying* ***`SeqE-Guard`*** *to multiple query paths $(S_t^{(1)})_{t \in \mathbb{N}}, (S_t^{(2)})_{t \in \mathbb{N}}, \dots$. Then, it holds for all $\mathbb{P} \in \mathcal{P}$: $\mathbb{P}(d_t^{(m)} \leq |S_t^{(m)} \cap I_1^{\mathbb{P}}| \text{ for all } t, m \in \mathbb{N}) \geq 1 - \alpha$.*

It should be noted that if we apply `SeqE-Guard` to multiple query paths, we must use the same e-values for every query path we consider.

Example 2 (Nontrivial true discovery bounds with weak signals)**.** *Let $\alpha = 0.05$ and suppose we test* 1000 *hypotheses using sequential e-values. Further suppose that* 300 *of those e-values equal* 0.9 *and* 700 *of those e-values equal* 1.1*. Due to the small e-values, it seems like there is no or at most very small evidence against the null hypotheses. Indeed, standard multiple testing procedures like the e-BH procedure for FDR control [37] would not make any rejections, regardless of the significance level $\alpha \in (0, 1]$. However, plugging these e-values into the* ***`SeqE-Guard`*** *algorithm yields a bound of $d_{1000} \geq 306$ for the query set $S_{1000} = \{i \leq 1000 : E_i > 1\}$, since $0.9^{300} 1.1^{395} \geq 20$. Consequently, the* ***`SeqE-Guard`*** *algorithm can claim with a probability of* 0.95 *that there are at least* 306 *false hypotheses among the* 700 *hypotheses with e-value greater than* 1*. This demonstrates the capability of the* ***`SeqE-Guard`*** *algorithm to detect small but frequent signals in the data.*

3.2 Improving existing methods

Existing online procedures with simultaneous true discovery guarantees were proposed by Katsevich and Ramdas [20] and Meah et al. [25]. Their setting involved observing one p-value for each hypothesis (as is standard in multiple testing) but in this section, we will show that these methods can be uniformly improved by `SeqE-Guard` by employing specific choices of the sequential e-values. Furthermore, we show that `SeqE-Guard` even uniformly improves some existing offline procedures.

3.2.1 Improving existing online procedures

We start with the `online-simple` procedure by Katsevich and Ramdas [20]. Suppose that p-values $P_1, P_2, \dots$ for the hypotheses $H_1, H_2, \dots$ are available such that P_i is measurable with respect to $\mathcal{F}_i$

Algorithm 1 `SeqE-Guard`: Online true discovery guarantee with sequential e-values

Input: Sequence of sequential e-values $E_1, E_2, \ldots$.
Output: Query sets $S_1 \subseteq S_2 \subseteq \ldots$ and true discovery bounds $d_1 \leq d_2 \leq \ldots$.

$d_0 = 0$
$S_0 = \emptyset$
$U = \emptyset$
$A = \emptyset$
for $t = 1, 2, \ldots$ **do**
    $S_t = S_{t-1}$
    $d_t = d_{t-1}$
    Statistician observes E_t and chooses whether index t should be included in S_t.
    if $t \in S_t$ **then**
        $A = A \cup \{t\}$
        if $\prod_{i \in A \cup U} E_i \geq 1/\alpha$ **then**
            $d_t = d_{t-1} + 1$
            $A = A \setminus \{\text{index of largest e-value in } A\}$
        end if
    else if $E_t < 1$ **then**
        $U = U \cup \{t\}$
    end if
    return S_t, d_t
end for

and let $\alpha_1, \alpha_2, \ldots$ be nonnegative thresholds such that α_i is measurable with respect to $\mathcal{F}_{i-1}$. It is assumed that the null p-values are valid conditional on the past, meaning $\mathbb{P}(P_i \leq x | \mathcal{F}_{i-1}) \leq x$ for all $i \in I_0^{\mathbb{P}}$, $x \in [0, 1]$. Katsevich and Ramdas [20] showed that

$$\boldsymbol{d}^{\text{os}}(S_t) = \left\lceil -ca + \sum_{i=1}^{t} \mathbb{1}\{P_i \leq \alpha_i\} - c\alpha_i \right\rceil \quad (t \in \mathbb{N}) \tag{13}$$

provides simultaneous true discovery guarantee over all sets $S_t = \{i \leq t : P_i \leq \alpha_i\}$, $t \in \mathbb{N}$, where $a > 0$ is some parameter and $c = \frac{\log(1/\alpha)}{a \log(1+\log(1/\alpha)/a)}$. On closer examination of their proof, one can observe that they proved their guarantee by implicitly showing that

$$E_i^{\text{os}} = \exp[\theta_c(\mathbb{1}\{P_i \leq \alpha_i\} - c\alpha_i)] \quad (i \in \mathbb{N}) \tag{14}$$

define sequential e-values, where $\theta_c = \log(1/\alpha)/(ca)$. We now propose to simply plug these sequential e-values into `SeqE-Guard`; this leads to Algorithm 2, which we will refer to as `closed online-simple` procedure in the following. To see this, note that

$$\prod_{i \in I} E_i^{\text{os}} \geq 1/\alpha \Leftrightarrow \sum_{i \in I} \mathbb{1}\{P_i \leq \alpha_i\} - c\alpha_i \geq \log(1/\alpha)/\theta_c = ca \quad (I \subseteq \mathbb{N}).$$

The superscript of the set A_t^c in Algorithm 2 is used to reflect the fact that A_t^c contains all indices of queried hypotheses that were excluded from the analysis until step t, and therefore can be seen as complement of the set A in Algorithm 1 with respect to S_t.

Algorithm 2 `Closed online-simple`

Input: Sequence of p-values $P_1, P_2, \ldots$ and sequence of (potentially data-dependent) individual significance levels $\alpha_1, \alpha_2, \ldots$.

Output: Query sets $S_1 \subseteq S_2 \subseteq \ldots$ and true discovery bounds $d_1 \leq d_2 \leq \ldots$.

$d_0 = 0$
$S_0 = \emptyset$
$A_0^c = \emptyset$
for $t = 1, 2, \ldots$ **do**
    if $P_t \leq \alpha_t$ **then**
        $S_t = S_{t-1} \cup \{t\}$
    else
        $S_t = S_{t-1}$
    end if
    if $\sum_{i \in \{1,\ldots,t\} \setminus A_{t-1}^c} \mathbb{1}\{P_i \leq \alpha_i\} - c\alpha_i \geq ca$ **then**
        $d_t = d_{t-1} + 1$
        $A_t^c = A_{t-1}^c \cup \{\text{index of smallest individual significance level in } S_t \setminus A_{t-1}^c\}$
    else
        $d_t = d_{t-1}$
        $A_t^c = A_{t-1}^c$
    end if
    return S_t, d_t
end for

Let A_t^c and d_t be defined as in `closed online-simple`, then

$$d_t \geq |A_{t-1}^c| + \left\lceil -ca + \sum_{i \in \{1,\ldots,t\} \setminus A_{t-1}^c} \mathbb{1}\{P_i \leq \alpha_i\} - c\alpha_i \right\rceil$$
$$= \left\lceil -ca + \sum_{i=1}^{t} \mathbb{1}\{P_i \leq \alpha_i\} - \sum_{i \in \{1,\ldots,t\} \setminus A_{t-1}^c} c\alpha_i \right\rceil \geq \boldsymbol{d}^{\text{os}}(S_t), \tag{15}$$

which shows that our `closed online-simple` method uniformly improves the `online-simple` procedure by Katsevich and Ramdas [20]. The improvement can be divided into two parts. First, the `closed online-simple` procedure is coherent, providing that $d_{t-1} \leq d_t$ for all $t \in \mathbb{N}$. Second, every time the bound d_t is increased by one, the summand $-c\alpha_i$ is excluded from the bound, where α_i is the smallest significance level with index in $S_t \setminus A_{t-1}^c$. This shows that the (online) closure principle and `SeqE-Guard` automatically adapt to the number of discoveries and thus the proportion of false hypotheses.

Note that E_i^{os} only takes two values. It takes $\exp[\theta_c(1 - c\alpha_i)]$ if $P_i \leq \alpha_i$ and $\exp[-\theta_c c\alpha_i]$ if $P_i > \alpha_i$, where $\mathbb{P}(P_i \leq \alpha_i | \mathcal{F}_{i-1}) \leq \alpha_i$ for all $\mathbb{P} \in H_i$. Hence, we have for all $\mathbb{P} \in H_i$,

$$\mathbb{E}_{\mathbb{P}}[E_i^{\text{os}} | \mathcal{F}_{i-1}] \leq \alpha_i \exp[\theta_c(1 - c\alpha_i)] + (1 - \alpha_i) \exp[-\theta_c c\alpha_i] =: u_i. \tag{16}$$

For example, $\alpha_i = \alpha = 0.1$ and $a = 1$ yield $u_i = 0.977$ (u_i is increasing in a; for $a = 3$ we obtain $u_i = 0.997$), which shows that E_i^{os} is not admissible and thus can be improved. A simple improvement can be obtained by plugging the e-value $\tilde{E}_i^{\text{os}} = E_i^{\text{os}}/u_i$ instead of E_i^{os} into `SeqE-Guard`. Doing this at every step, we obtain a further improvement; we call this the `admissible online-simple` method.

Proposition 3.3. *By plugging in the sequential e-values* $(E_i^{os})_{i \in \mathbb{N}}$ (14) *into the* ***`SeqE-Guard`*** *algorithm, we obtain multiple uniform improvements (that can be applied together) over the* ***`online-simple`*** *method by Katsevich and Ramdas [20]:*

1. *The lower bound d_t is nondecreasing in t (`coherent online-simple`).*

2. *Every time the bound d_t is increased by one, the summand $-c\alpha_i$ is excluded from the bound, where α_i is the smallest threshold with index in $S_t \setminus A_{t-1}^c$ (`closed online-simple`).*

Furthermore, since the expected value of E_i^{os} is strictly smaller than 1 under H_0, an additional improvement can be obtained:

3. *The sequential e-values E_i^{os}, $i \in \mathbb{N}$, can be replaced by $\tilde{E}_i^{os} = E_i^{os}/u_i$, where $u_i < 1$ is given by (16) (`admissible online-simple`).*

Katsevich and Ramdas [20] introduced one further online procedure with simultaneous true discovery guarantee, the `online-adaptive` method. A uniform improvement can be obtained in the exact same manner as for the `online-simple` method above; we present this in Supplementary Material S.3.1. Inspired by the methods of Katsevich and Ramdas [20], two further online procedures with simultaneous true discovery guarantee were proposed by Meah et al. [25]. The first procedure is obtained by taking a union of `online-simple` bounds for different parameters a. The second procedure exploits Freedman's inequality [10] and a union bound. In Supplementary Material S.3.2, we show how both these recent methods can be uniformly improved by our e-value based approach. The proposed improvements are technically not instances of the `SeqE-Guard` algorithm, but can be obtained by the union of `SeqE-Guard` bounds (or using the average of multiple test martingales). In Table 1 we provide the concrete sequential e-values yielding the mentioned improvements.

3.2.2 Improving existing offline procedures

Interestingly, even though the `SeqE-Guard` algorithm was specifically designed for the *online* multiple testing setting, it also allows to improve some *offline* methods. Katsevich and Ramdas [20] introduced true discovery procedures for pre-ordered and interactive paths. That means, even though all data are available at the beginning, the hypotheses are ordered based on some side information or orthogonal data. For instance, this includes simultaneous true discovery bounds for the knockoff framework [3, 6].

These offline procedures are based on the same idea as the `online-simple` and `online-adaptive` method (see Lemma 1 in Katsevich and Ramdas [20]). Given ordered p-values $P_{(1)}, P_{(2)}, \ldots$ and nonnegative thresholds $\alpha_1, \alpha_2, \ldots$, they proposed the general bounds

$$\boldsymbol{d}^{\mathrm{KR}}(S_t) = -c_\alpha a + \sum_{i \le t} \mathbb{1}\{P_{(i)} \le \alpha_i\} - c_\alpha h_i(P_{(i)}) \quad (t \in \mathbb{N})$$

for the query path $S_t = \{i \le t : P_{(i)} \le \alpha_i\}$, $t \in \mathbb{N}$, where $(h_i)_{i\in\mathbb{N}}$ are functions on $[0,1]$, $a > 1$ some regularization parameter and c a constant. They implicitly proved that $\boldsymbol{d}^{\mathrm{KR}}$ provides simultaneous true discovery guarantee over all S_t, $t \in \mathbb{N}$, if $E_{(1)}^{\mathrm{KR}}, E_{(2)}^{\mathrm{KR}}, \ldots$ define sequential e-values, where

$$E_{(i)}^{\mathrm{KR}} = \exp\left[\frac{\log(1/\alpha)}{ac}\left(\mathbb{1}\{P_{(i)} \le \alpha_i\} - c h_i(P_{(i)})\right)\right] \quad (i \in \mathbb{N}), \tag{17}$$

and gave specific choices of $(h_i)_{i\in\mathbb{N}}$, $(\alpha_i)_{i\in\mathbb{N}}$ and c that satisfy this requirement. It is apparent that this general approach recovers the `online-simple` method for $(i) = i$ and $h_i(P_{(i)}) = \alpha_i$. Furthermore, it can easily be concluded that we can improve these general bounds $\boldsymbol{d}^{\mathrm{KR}}$ by plugging the e-values $E_{(1)}^{\mathrm{KR}}, E_{(2)}^{\mathrm{KR}}, \ldots$ into `SeqE-Guard`, as we illustrated with the `online-simple` method in Section 3.2.1. Since the main focus of this paper is on the online setting, we will not discuss each of the improvements individually.

Existing method	Sequential e-values	Algorithm
`online-simple` [20]	$E_i^{\text{os}} = \exp[\theta_c(\mathbb{1}\{P_i \leq \alpha_i\} - c\alpha_i)]$ $a > 0$ $c = \frac{\log(1/\alpha)}{a\log(1+\log(1/\alpha)/a)}$ $\theta_c = \log(1/\alpha)/(ca)$	`SeqE-Guard`
`online-adaptive` [20]	$E_i^{\text{ad}} = \exp[\theta_c(\mathbb{1}\{P_i \leq \alpha_i\} - c\frac{\alpha_i}{1-\lambda_i}\mathbb{1}\{P_i > \lambda_i\})]$ $\lambda_i \in [\alpha_i, 1)$ $a > 0$ $c = \frac{\log(1/\alpha)}{a\log(1+(1-\alpha^{B/a})/B)}$ $B = \sup_{i\in\mathbb{N}} \frac{\alpha_i}{1-\lambda_i}$ $\theta_c = \log(1/\alpha)/(ca)$	`SeqE-Guard`
`u-online-simple` [25]	$E_i^{\text{os},a} = \exp[\theta_{c_a}(\mathbb{1}\{P_i \leq \alpha_i\} - c_a\alpha_i)]$ $c_a = \frac{\log(1/\alpha(a))}{a\log(1+\log(1/\alpha(a))/a)}$ $\alpha(a) = \frac{6\alpha}{a^2\pi^2}$ $\theta_{c_a} = \log(1/\alpha(a))/(c_a a)$	`SeqE-Guard` + union over $a \in \mathbb{N}$
`u-online-Freedman` [25]	$E_i^{\text{Freed},a} = \exp[\lambda_a(\mathbb{1}\{P_i \leq \alpha_i\} - \alpha_i)$ $-\psi(\lambda_a)\alpha_i(1-\alpha_i)]$ $\lambda_a = \log\left(1 + \frac{\kappa_a}{a}\right)$ $\psi(\lambda_a) = \exp(\lambda_a) - \lambda_a - 1$ $\kappa_a = \sqrt{2a\log(1/\alpha(a))} + \frac{\log(1/\alpha(a))}{2}$ $\alpha(a) = \alpha\left(\frac{6}{\max(2\log_2(a),1)^2(\pi^2+6)}\right)$	`SeqE-Guard` + union over $a = 2^{j/2}$, $j \in \mathbb{N}\cup\{0\}$

Table 1: List of online procedures that can be uniformly improved by our `SeqE-Guard` algorithm. The left column shows the name of the existing method proposed in past work. The middle column gives the sequential e-value, which if plugged into the algorithm in the right column, yields a uniform improvement of the existing method in the left column. To the best of our knowledge this list includes all existing nontrivial online true discovery procedures. Thus, we claim that our approach uniformly improves all existing procedures. It should be noted that the shown sequential e-values may be further improved by ensuring that their expected value equals exactly 1 in case of uniformly distributed p-values (this can be accomplished by dividing the shown e-values by their expected value but we omit it here to keep the expressions short).

3.3 Adaptively hedged GRO e-values

The most common strategy to calculate e-values in practice is based on variants of the GRO-criterion [30, 17, 38], which dates back to the Kelly criterion [21, 5]. Here, we assume that each null hypothesis H_i comes with an alternative H_i^A. Suppose H_i^A contains a single distribution $\mathbb{Q}_i$. Then the growth rate optimal (GRO) e-value E_i^{GRO} is defined as the e-value E_i that maximizes the growth rate under $\mathbb{Q}_i$, given by $\mathbb{E}_{\mathbb{Q}_i}[\log(E_i)]$, over all e-values for H_i. If H_i^A is composite, a prior over its distributions can be used, defining the GRO e-value via the corresponding mixture distribution. If the null hypothesis is simple, the GRO e-value is given by the likelihood ratio (LR) of the alternative over the null distribution [30]. If H_i is composite, the GRO e-value takes the form of a LR of the alternative against a specific (sub-) distribution [17, 23]. The GRO e-value is particularly powerful when many e-values are combined by multiplication [30] and therefore seems to be a reasonable choice for our `SeqE-Guard` algorithm. Furthermore, since the growth rate is the standard measure of performance for e-values [29], there might be applications where we do not have access to the data to calculate our own e-value, but just get a GRO e-value for each individual hypothesis. For example, this could be the case in meta-analyses, where each study just reported the GRO e-value. In the following, we will discuss how the GRO concept transfers to online true discovery guarantee based on sequential

e-values.

A naive approach would be to just plugin the GRO e-values into `SeqE-Guard`. However, the GRO e-values only maximize the growth rate under their alternatives. This makes sense when testing a single hypothesis. In our setting, the product of GRO e-values would only maximize the growth rate of $W_t = \prod_{i=1}^t E_i$ if all hypotheses are false — a very unlikely scenario. Indeed, GRO e-values can be small if the null hypothesis is true, so directly using GRO e-values in `SeqE-Guard` can lead to low power. Hence, when multiplying e-values for different hypotheses one needs to incorporate the possibility that a hypothesis is true. One approach is to hedge the GRO e-values by defining

$$\tilde{E}_i^{\mathrm{GRO}} = 1 - \lambda_i + \lambda_i E_i^{\mathrm{GRO}},$$

where $\lambda_i \in [0,1]$ is measurable with respect to $\mathcal{F}_{i-1}$. To see that $\tilde{E}_1^{\mathrm{GRO}}, \tilde{E}_2^{\mathrm{GRO}}, \dots$ indeed define sequential e-values if the GRO e-values are sequential, just note that $\mathbb{E}_P[\tilde{E}_i^{\mathrm{GRO}} | \mathcal{F}_{i-1}] = 1 - \lambda_i + \lambda_i \mathbb{E}_{\mathbb{P}}[E_i^{\mathrm{GRO}} | \mathcal{F}_{i-1}] \leq 1$ for all $\mathbb{P} \in H_i$.

In order to derive a reasonable choice for λ_i, we consider a specific Bayesian two-groups model. Suppose $H_i = \{\mathbb{P}_i\}$ and $H_i^A = \{\mathbb{Q}_i\}$ are both simple hypotheses and which of these hypotheses is true is random, where τ_i gives the probability that the alternative hypothesis H_i^A is true. A reasonable approach would be to choose the e-value E_i for H_i that maximizes the growth rate under the true distribution $\mathbb{E}_{(1-\tau_i)\mathbb{P}_i + \tau_i \mathbb{Q}_i}[\log(E_i)]$.

Proposition 3.4. *Suppose $\mathbb{Q}_i$ is absolutely continuous with respect to $\mathbb{P}_i$. Then the e-value E_i that maximizes the growth rate under the true distribution $(1-\tau_i)\mathbb{P}_i + \tau_i \mathbb{Q}_i$ is given by $1 - \tau_i + \tau_i E_i^{\mathrm{GRO}}$, where E_i^{GRO} maximizes the growth rate under the alternative $\mathbb{Q}_i$.*

Proof. Due to Shafer [30], the e-value maximizing the growth rate under $(1-\tau_i)\mathbb{P}_i + \tau_i \mathbb{Q}_i$ is given by the likelihood ratio

$$\frac{d[(1-\tau_i)\mathbb{P}_i + \tau_i \mathbb{Q}_i]}{d\mathbb{P}_i} = 1 - \tau_i + \tau_i \frac{d\mathbb{Q}_i}{d\mathbb{P}_i} = 1 - \tau_i + \tau_i E_i^{\mathrm{GRO}},$$

where $\frac{d\mathbb{Q}_i}{d\mathbb{P}_i}$ denotes the Radon-Nikodym derivative. □

Proposition 3.4 shows that in the case of simple null and alternative hypotheses the optimal e-value under the true distribution is the same as if we maximize the growth rate under the alternative and then hedge the resulting e-value according to the probability that the alternative is true. We think it is possible that something analogous holds for composite null hypotheses as well. However, even if this is not the case, it seems to be a reasonable strategy in general.

Therefore, our approach is to specify an estimate $\hat{\tau}_i$ for the probability τ_i that H_i is false, where $\hat{\tau}_i$ can either depend on prior information or the past data, and then set $\lambda_i = \hat{\tau}_i$. We propose to set

$$\hat{\tau}_i = \frac{1/2 + \sum_{j=1}^{i-1} \mathbb{1}\{E_j^{\mathrm{GRO}} > 1\}}{i} \qquad (i \in \mathbb{N}), \tag{18}$$

if no prior information is available. The idea is that $E_j^{\mathrm{GRO}} > 1$ indicates data favoring the alternative over the null distribution, since the GRO e-value can be interpreted as a (generalized) LR.

We illustrate the behavior of $\hat{\tau}_i$ defined in (18) for different scenarios in Figure 3. We consider the simulation setup described in Section 4, but with $n = 100$ and proportion of false hypotheses $\pi_A \in \{0.2, 0.5, 0.8\}$. If the signal is weak, $\hat{\tau}_i$ lies between the true proportion and 0.5 and if the signal is strong, the estimate $\hat{\tau}_i$ is close to the true proportion. We think that this behavior is desirable, since a tendency towards 0.5 is not that hurtful if the alternative and null distribution are close as the GRO e-values have a small variance in this case.

Hedging e-values before multiplying them does not only apply to GRO e-values. Such strategies for merging sequential e-values have been used by many preceding authors like Waudby-Smith and Ramdas [38], Vovk and Wang [35]. However, the above argumentation provides a reasonable choice for the parameter λ_i in our setting. We analyze this approach experimentally in Section 4.2.

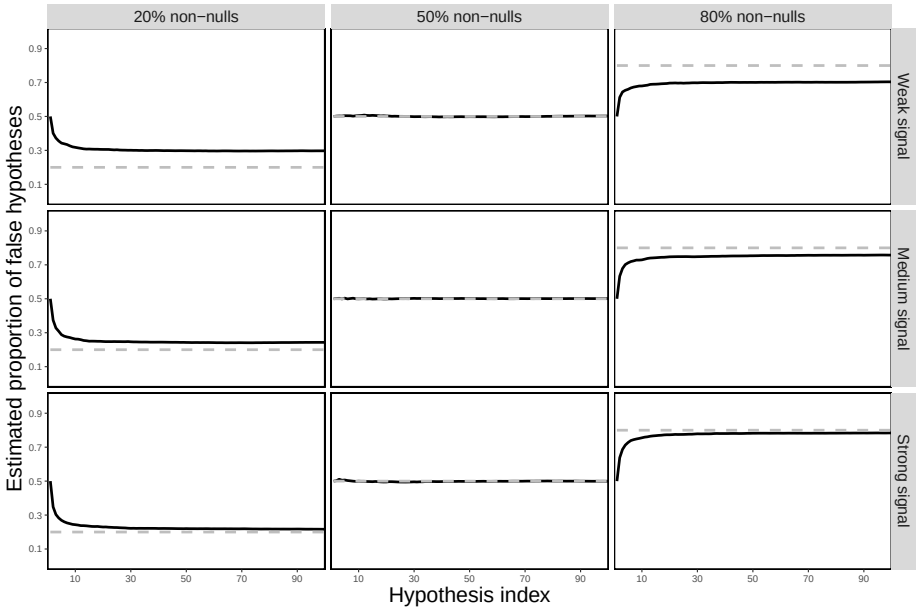


Figure 3: Comparing $\hat{\tau}_i$ (solid line) defined in (18) to the true τ_i (dashed line) for different scenarios. The simulation design is described in Section 4. The estimate $\hat{\tau}_i$ has a tendency to bias towards 50% but it estimates the proportion of false hypotheses very well when the signal is strong.

4 Simulations

In this section we numerically calculate the true discovery proportion (TDP) bound, which is defined as the true discovery bound for S_t divided by the size of S_t. We compare TDP bounds obtained by applying `SeqE-Guard` to the different sequential e-values proposed in the previous sections. In Subsection 4.1, we compare the `online-simple` method by Katsevich and Ramdas [20] with its uniform improvement. In Subsection 4.2, we demonstrate how hedging GRO e-values improve the true discovery bound. In Subsection 4.3, we compare the proposed e-values to decide which is best suited for practice. An application to real data can be found in Supplementary Material S.6.

We consider the same simulation setup in all subsections. We sequentially test $n = 1000$ null hypotheses H_i, $i \in \{1, \dots, n\}$, of the from $H_i : X_i \sim \mathcal{N}(0, 1)$ against the alternative $H_i^A : X_i \sim \mathcal{N}(\mu_A, 1)$ for some $\mu_A > 0$, where $X_1, \dots, X_n$ are independent data points or test statistics. The probability that the alternative hypothesis is true is set by a parameter $\pi_A \in (0, 1)$ and the desired guarantee is set to $\alpha = 0.1$. For all comparisons we consider a grid of simulation parameters $\mu_A \in \{2, 3, 4\}$ and $\pi_A \in \{0.1, 0.3, 0.5\}$, where we refer to $\mu_A = 2$ as weak signal, $\mu_A = 3$ as medium signal and $\mu_A = 4$ as strong signal. The p-values are calculated by $\Phi(-X_i)$, where Φ is the CDF of a normal distribution. The raw GRO e-values are given by the likelihood ratio $E_i^{\mathrm{GRO}} = p_{\mu_A}(X_i)/p_0(X_i)$, where p_{μ_A} and p_0 are the densities of a normal distribution with variance 1 and mean μ_A and 0, respectively. The query sets S_t, $t \in \{1, \dots, t\}$, are defined as $S_t = \{i \in \{1, \dots, t\} : P_i \leq \alpha\}$. All of the results in the following are obtained by averaging over 1000 independent trials.

4.1 Comparing the `online-simple` method [20] with its improvements

In Section 3.2 we showed that the `online-simple method` by Katsevich and Ramdas [20] can be uniformly improved by the `closed online-simple` procedure (Algorithm 2). We also showed that this closed procedure can be further uniformly improved by the `admissible online-simple`

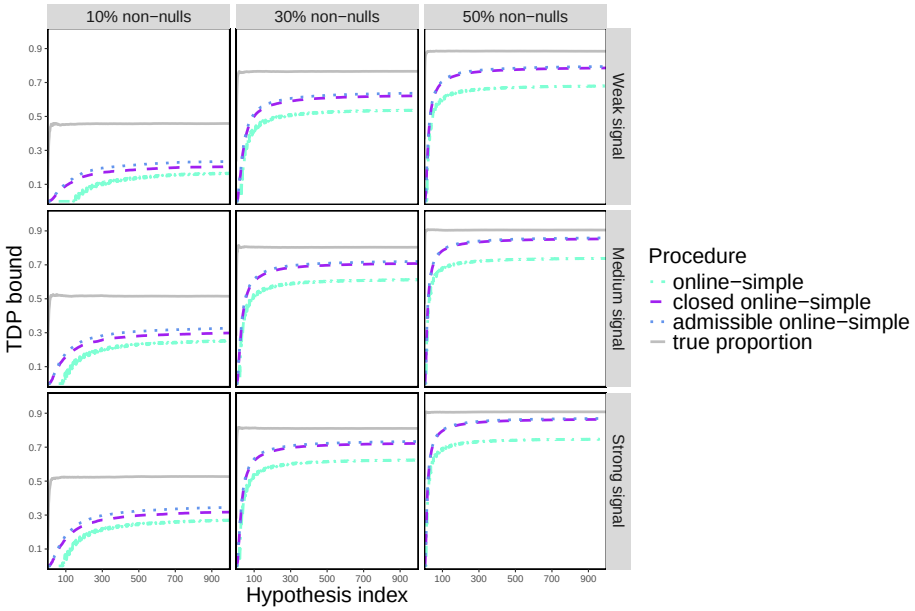


Figure 4: True discovery proportion bounds obtained by the `online-simple` method [20], the `closed online-simple` method and the `admissible online-simple` method. The `closed online-simple` method leads to substantially larger bounds than the original `online-simple` method. An additional improvement can be obtained by the `admissible online-simple` method, which is closest to the true proportion (top line) in all figures.

procedure, which ensures that the expected value of each sequential e-value is exactly one. In this section, we aim to quantify the gain in power for making true discoveries by using these improvements instead of the `online-simple` method. Although Katsevich and Ramdas [20] proposed $a = 1$ as default parameter, we found $a = 3$ to perform better which is why we use it here.

The results are illustrated in Figure 4. It can be seen that the `closed online-simple` procedure leads to a substantial improvement of the `online-simple` procedure in all cases. Of the queried hypotheses, the former approximately identifies $10\% - 20\%$ more as false. The additional improvement of the admissible online-simple method is quite small in this case, however, it is potentially larger for smaller parameters a (see Section 3.2).

In Section S.3.2 we also provide simulations for our improvements over the methods proposed by Meah et al. [25]. The behavior of both, the initially proposed methods and our improvements, is quite similar as in the `online-simple` case.

4.2 GRO e-values

In Section 3.3 we argued that the raw GRO e-values should be hedged to account for the probability that a null hypothesis is true. Now we compare the true discovery bounds obtained by applying `SeqE-Guard` to raw and hedged GRO e-values. For the hedged GRO e-values we chose the predictable parameter proposed in (18). The results are illustrated in Figure 5, showing that GRO e-values lead to very low bounds which can be increased substantially by hedging.

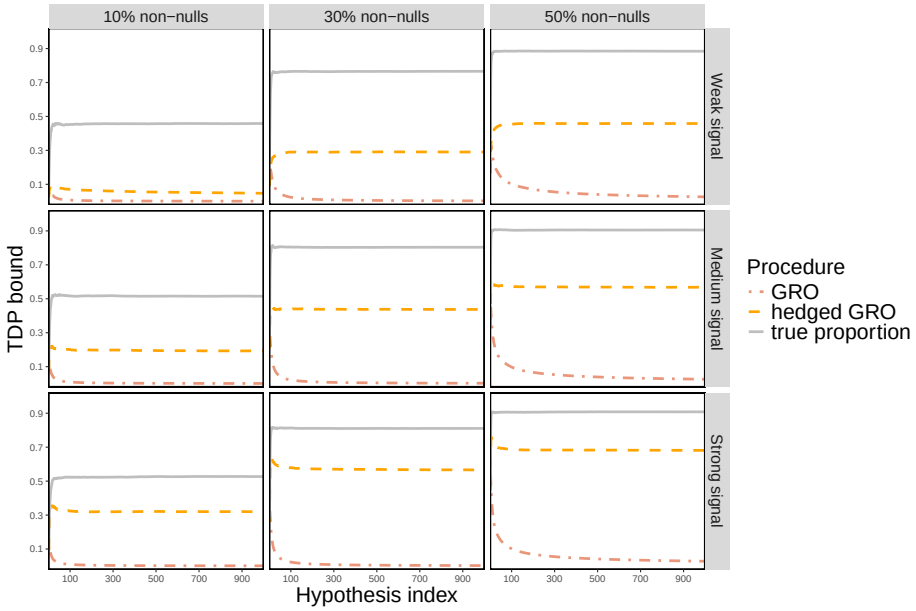


Figure 5: True discovery proportion bounds obtained by applying `SeqE-Guard` to GRO e-values and hedged GRO e-values. The hedged GRO e-values improve the GRO e-values substantially.

4.3 Which sequential e-values should we choose?

In this section, we compare the `SeqE-Guard` procedure when applied with the best versions of the proposed sequential e-values to derive recommendations for practice. More precisely, we compare the `admissible online-simple` method with the hedged GRO e-value.

The results are depicted in Figure 6. The procedures perform quite differently in the various settings. When the proportion of false hypotheses is small, the `SeqE-Guard` algorithm performs best with hedged GRO e-values, particularly, if the signal is strong. However, if the proportion of false hypotheses is large, the `admissible online-simple` method clearly outperforms the `SeqE-Guard` with hedged GRO e-values. Hence, if we expect sparse but strong signal, applying `SeqE-Guard` with hedged GRO e-values is the best choice. In contrast, for dense but weak signals, the `admissible online-simple` method should be preferred.

Furthermore, it should be noted that the only reasonable query path for the `online-simple` method and its improvements is given by $S_t = \{i \leq t : P_i \leq \alpha_i\}$ (which we use in the simulations for $\alpha_i = \alpha$). To see this, note that all e-values E_i^{os}, $i \leq t$, with $i \notin S_t$ are smaller than 1 and therefore their inclusion in S_t would not increase the lower bound for the number of true discoveries. Furthermore, excluding e-values E_i^{os}, $i \leq t$, with $P_i \leq \alpha_i$ would be difficult to communicate, as those e-values reached there maximum possible value. Hence, hedged GRO e-values are better suited for an exploratory analysis where the scientist might be interested in several different query paths (see Corollary 3.2). For example, `SeqE-Guard` with hedged GRO e-values can provide a (nontrivial) query path with online FWER control (by including $t \in S_t$ iff this implies $d_t = d_{t-1} + 1$), while simultaneously providing (nontrivial) real-time lower bounds for the number of false hypotheses among all hypotheses with e-values greater than, say, 2. For such an exploratory proceeding, we would recommend the hedged GRO e-values, since they showed good performance (Figure 5) without adapting to the query path at all. In Supplementary Material S.5, we show how GRO e-values could be adapted to the chosen query path using a boosting approach [37] that uniformly improves the bounds for that specific query path (but may reduces the bounds for other query paths). In Supplementary Material S.4 we show how one could investigate multiple query paths when only

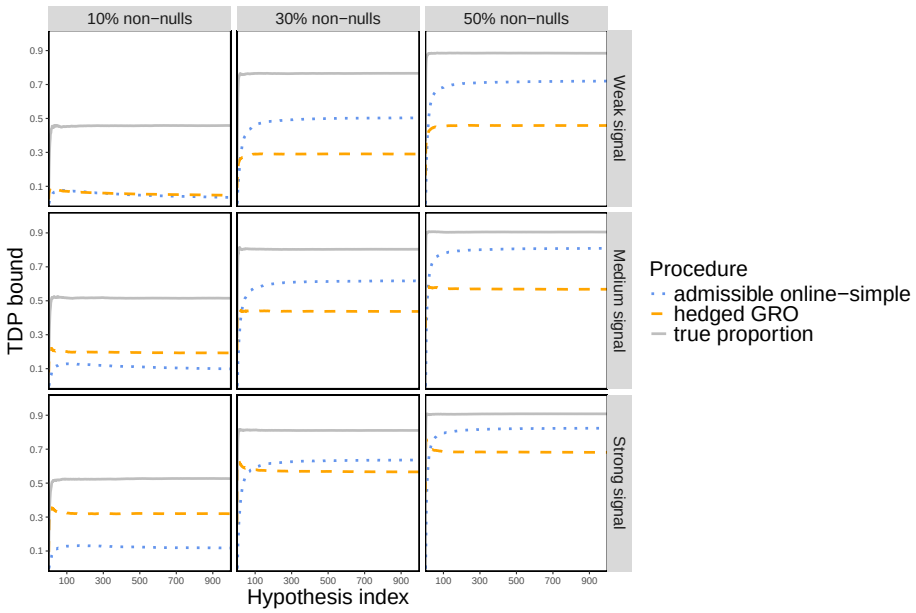


Figure 6: True discovery proportion bounds obtained by the `admissible online-simple` method and applying `SeqE-Guard` to hedged GRO e-values. If the proportion of false hypotheses is small, the hedged GRO e-values perform well, particularly if the signal is strong. If the proportion of false hypotheses is large, the `admissible online-simple` method should be preferred. However, the `admissible online-simple` method adapts to the specific query path, while the hedged GRO e-values are better suited for exploring multiple query paths.

p-values are available, using other p-to-e calibrators than the simple binary calibrator applied for the `online-simple` method.

5 Discussion

In this paper, we proposed a new closed testing based online true discovery procedure for sequential e-values and derived a general short-cut that only requires one calculation per hypothesis. Although the `SeqE-Guard` algorithm is restricted to sequential e-values, it is a general procedure for the task of online true discovery guarantee, since there are many different ways to construct sequential e-values. In particular, it yields uniform improvements of the state-of-the-art methods by Katsevich and Ramdas [20] and Meah et al. [25], although they were not explicitly constructed using e-values.

From a theoretical point of view this paper gives new insights about the role of e-values in multiple testing by showing that every admissible coherent online true discovery procedure must be based on sequential e-values. From a practical point of view, we constructed a powerful and flexible multiple testing procedure, which allows to observe hypotheses one-by-one over time and make fully data-adaptive decisions about the hypotheses and stopping time while bounding the number of true discoveries or equivalently, the false discovery proportion. On the way, we introduced a new idea for hedging of GRO e-values in multiple testing.

Interestingly, even though we constructed the `SeqE-guard` algorithm primarily for the online setting, it also allows to uniformly improve some existing offline procedures. Future work could examine the implications of our results for the offline setting in more detail.

Acknowledgments

The authors thank Etienne Roquain, two anonymous reviewers, an associate editor and a joint editor for their helpful comments. LF acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 281474342/GRK2224/2. AR was funded by NSF grant DMS-2310718.

References

[1] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526 (7571):68, 2015.

[2] 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*, 76(4):771–794, 2014.

[3] Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[4] 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: Statistical Methodology*, 57(1):289–300, 1995.

[5] Leo Breiman. Optimal gambling systems for favourable games. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 65–78, 1961.

[6] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.

[7] Jean Feng, Scott Emerson, and Noah Simon. Approval policies for modifications to machine learning-based software as a medical device: A study of bio-creep. *Biometrics*, 77(1):31–44, 2021.

[8] Lasse Fischer, Marta Bofill Roig, and Werner Brannath. The online closure principle. *The Annals of Statistics*, 52(2):817–841, 2024.

[9] Dean P Foster and Robert A Stine. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(2): 429–444, 2008.

[10] David A Freedman. On tail probabilities for martingales. *The Annals of Probability*, pages 100–118, 1975.

[11] K Ruben Gabriel. Simultaneous test procedures–some theory of multiple comparisons. *The Annals of Mathematical Statistics*, 40(1):224–250, 1969.

[12] Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035–1061, 2004.

[13] Christopher R Genovese and Larry Wasserman. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476):1408–1417, 2006.

[14] Jelle J Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

[15] Jelle J Goeman, Rosa J Meijer, Thijmen JP Krebs, and Aldo Solari. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106(4): 841–856, 2019.

[16] Jelle J Goeman, Jesse Hemerik, and Aldo Solari. Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics*, 49(2):1218–1238, 2021.

[17] Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology (with discussion)*, 2024.

[18] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[19] 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.

[20] Eugene Katsevich and Aaditya Ramdas. Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *The Annals of Statistics*, 48 (6):3465–3487, 2020.

[21] John L Kelly. A new interpretation of information rate. *The Bell System Technical Journal*, 35 (4):917–926, 1956.

[22] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1168–1176, 2013.

[23] Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *The Annals of Statistics*, 2025.

[24] Ruth Marcus, Peritz Eric, and K Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[25] Iqraa Meah, Gilles Blanchard, and Etienne Roquain. False discovery proportion envelopes with m-consistency. *Journal of Machine Learning Research*, 25(270):1–52, 2024.

[26] Violeta Muñoz-Fuentes, Pilar Cacheiro, Terrence F Meehan, Juan Antonio Aguilar-Pimentel, Steve DM Brown, Ann M Flenniken, Paul Flicek, Antonella Galli, Hamed Haseli Mashhadi, Martin Hrabě de Angelis, et al. The international mouse phenotyping consortium (IMPC): a functional catalogue of the mammalian genome that informs conservation. *Conservation genetics*, 19(4):995–1005, 2018.

[27] Aaditya Ramdas, Fanny Yang, Martin J Wainwright, and Michael I Jordan. Online control of the false discovery rate with decaying memory. *Advances in Neural Information Processing Systems*, 30, 2017.

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

[29] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

[30] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society (with discussion)*, 184(2):407–431, 2021.

[31] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 2011.

[32] Eckart Sonnemann. *Allgemeine Lösungen multipler Testprobleme*. Universität Bern. Institut für Mathematische Statistik und Versicherungslehre, 1982.

[33] Jinjin Tian, Xu Chen, Eugene Katsevich, Jelle Goeman, and Aaditya Ramdas. Large-scale simultaneous inference under dependence. *Scandinavian Journal of Statistics*, 50(2):750–796, 2023.

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

[35] Vladimir Vovk and Ruodu Wang. Merging sequential e-values via martingales. *Electronic Journal of Statistics*, 18(1):1185–1205, 2024.

[36] A Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117–186, 1945.

[37] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022.

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

[39] Ziyu Xu and Aaditya Ramdas. Dynamic algorithms for online multiple testing. In *Mathematical and Scientific Machine Learning*, pages 955–986. PMLR, 2022.

[40] 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*, pages 3806–3815. PMLR, 2020.

[41] Tijana Zrnic, Aaditya Ramdas, and Michael I Jordan. Asynchronous online testing of multiple hypotheses. *Journal of Machine Learning Research*, 22(33):1–39, 2021.

Supplementary material for "Admissible online closed testing must employ e-values"

S.1 Related literature

Our work mixes ingredients from different subfields of sequential and multiple hypothesis testing.

The e-value has recently emerged as a fundamental concept in composite hypothesis testing and underlying a universal approach to anytime-valid inference [42, 30, 36, 13, 25], but the roots can be traced back to the works of Ville [35], Wald [40] and Robbins [3]. A recent overview of the e-value literature is given by Ramdas et al. [26] and Ramdas and Wang [23, Chapter 1].

Interest in e-values has grown rapidly in recent years, including in particular multiple testing with e-values. Wang and Ramdas [41] introduced and analyzed the e-BH procedure, an e-value variant of the popular Benjamini-Hochberg (BH) procedure [2] for FDR control. Vovk and Wang [36] explored the possibility of combining several e-values by averaging and multiplication. They also used this to derive multiple testing procedures with familywise error rate (FWER) control by applying the closure principle [20] with these combination rules. The FWER is a strict error criterion defined as the probability of rejecting any true null hypothesis. Vovk and Wang [37, 38] extended these ideas to obtain procedures with a true discovery guarantee. All the aforementioned approaches consider classical *offline* multiple testing.

Online multiple testing initially focused on procedures for p-values [6, 1, 16, 24]. An overview of this literature was recently provided by Robertson et al. [28]. Xu and Ramdas [43] is the sole paper to consider online multiple testing with e-values, focusing on FDR control for dependent e-values.

A related line of work investigates simultaneous true discovery guarantees by closed testing, mostly focusing on offline settings with p-values. The closure principle was initially proposed and analyzed for FWER control [20, 31, 29]. However, Goeman and Solari [10] noted that the same principle can be applied to obtain simultaneous true discovery bounds, a more general and less conservative task than FWER control, although a similar approach was proposed earlier by Genovese and Wasserman [8, 9]. Many works have since built on these [11, 34, 14, 19]. Importantly, Goeman et al. [12] proved that all admissible procedures for bounding the true discovery proportion must employ closed testing.

The recent work of Fischer et al. [5] showed how the closure principle can also be used for *online* multiple testing. However, their investigation of admissibility and construction of concrete procedures is restricted to FWER control. The current work extends their ideas to lower bounds on the true discovery proportion.

Another related work to ours is by Katsevich and Ramdas [18]. They proposed various p-value based true discovery procedures for structured, knockoff and also online settings exploiting martingale techniques. Meah et al. [21] modified and improved some of their methods with a focus on m-consistency, a property that relates true discovery procedures to FDR. Our work uniformly improves the methods by Katsevich and Ramdas [18] and Meah et al. [21] for the online setting.

In Table S.1, we give a brief comparison of our theoretical results to the related works. Our paper is the first connecting online multiple testing to anytime-valid testing of a single hypothesis. Furthermore, our setting generalizes the ones considered by Goeman et al. [12] (since the online setting generalizes the offline setting) and Fischer et al. [5] (since true discovery guarantee generalizes FWER control). Consequently, our results immediately apply in their settings, but their results do not imply ours.

Paper	Key theoretical messages
Goeman et al. [12]	• Admissible closed testing is sufficient for admissibility of monotone stacks of true discovery procedures • All admissible true discovery procedures are closed procedures
Fischer et al. [5]	• Increasing families of online intersection tests are sufficient to obtain online closed procedures • All admissible FWER controlling online procedures are online closed procedures
Ramdas et al. [25]	• Sufficient conditions for admissibility of anytime-valid tests based on test martingales • All admissible anytime-valid tests must employ test martingales
This paper	• Online closed procedures must employ anytime-valid tests for the intersection hypotheses • All admissible coherent online true discovery procedures are online closed procedures.

Table S.1: Comparison of the key messages of closely related theoretical works. Our main claim, the first shown above, connects the work by Ramdas et al. [25] with the online multiple testing literature [5, 18, 21] and thus allows to apply the powerful anytime-valid testing concept of test martingales to online multiple testing. Furthermore, our second claim implies and unifies the second claims of Goeman et al. [12] and Fischer et al. [5], since the online setting generalizes the offline setting and true discovery guarantee generalizes FWER control (and all admissible procedures are coherent in these settings).

S.2 Simple example: Lower bounds for the number of false hypotheses

A simple but interesting special case of the general online discovery process (see Figure 1) occurs if we choose $S = \{1, \dots, t\}$ at each time $t \in \mathbb{N}$. That means we observe a stream of hypotheses $H_1, H_2, \dots$ with sequential e-values $E_1, E_2, \dots$ and want to provide a real-time lower bound $\boldsymbol{d}(\{1, \dots, t\}) = d_t$ for the number of false hypotheses among $H_1, \dots, H_t$ which holds true with high probability simultaneously over all times t:

$$\mathbb{P}(d_t \leq |\{i \leq t : H_i \text{ false}\}| \text{ for all } t \in \mathbb{N}) \geq 1 - \alpha \quad \text{ for some } \alpha \in [0, 1]. \tag{S.1}$$

In this case, the `SeqE-Guard` algorithm consists at each time t of two simple steps: (1) multiply the e-values up to step t; (2) if the product is greater or equal than $1/\alpha$, increase the lower bound by 1 and exclude the largest e-value from the future analysis. More precisely, set $d_0 = 0$ and $A = \emptyset$, and then do for $t = 1, 2, \dots$:

1. Set $A = A \cup \{t\}$ and calculate $\Pi = \prod_{i \in A} E_i$.
2. If $\Pi \geq 1/\alpha$, then update $d_t = d_{t-1} + 1$ and $A = A \setminus \{\text{index of largest e-value in } A\}$; otherwise, set $d_t = d_{t-1}$.

For example, suppose $\alpha = 0.05$ and the first five e-values are

$$E_1 = 5, \quad E_2 = 4, \quad E_3 = 0.8, \quad E_4 = 0.5, \quad E_5 = 14.$$

At time $t = 2$ the product (of E_1 and E_2) equals 20 and therefore we can set $d_2 = 1$ and then exclude E_1 from the future analysis. Then, at time $t = 5$, the product (of E_2, E_3, E_4 and E_5) is again greater

than 20 and therefore we can increase the lower bound and set $d_5 = 2$. Hence, in this case we can confidently claim (with probability 0.95) that there is at least one false hypothesis among H_1 and H_2 and at least two false hypotheses among $H_1, \ldots, H_5$. This claim remains valid regardless of how many hypotheses are tested in the future and what the e-values for these hypotheses look like.

Of course the claims above are too imprecise for many applications, as they only state that two of the five hypotheses are false, but not which ones. However, this is only a simple example. In general, the users of our `SeqE-Guard` algorithm can specify (based on the data) any subset of hypotheses in which they are interested and the `SeqE-Guard` algorithm will provide a lower bound for the number of false hypotheses in the subset that is valid simultaneously over all times and possible subsets. For example, a user might only be interested in the number of false hypotheses among H_1, H_2 and H_5, since due to their small e-values H_3 and H_4 are unlikely to be false anyway. The `SeqE-Guard` algorithm would then still provide a lower bound of 2 for this subset, since excluding e-values smaller than 1 from the query set won't decrease the bound (see Algorithm 1). This bound would be much more informative than the same lower bound for all five hypotheses, although we cannot make claims about the truth of individual hypotheses either. However, note that this is the same for FDR procedures, as they just state that the expected proportion of true hypotheses among the rejected ones is below α.

S.3 Uniform improvements of existing methods

S.3.1 Uniform improvement of the `online-adaptive` method by Katsevich and Ramdas [18]

Let p-values $P_1, P_2, \ldots$ and significance levels $\alpha_1, \alpha_2, \ldots$ be defined as for the `online-simple` algorithm (see Section 3.2), and the null p-values be valid conditional on the past. Furthermore, let $(\lambda_i)_{i\in\mathbb{N}}$ be additional parameters such that $\lambda_i \in [\alpha_i, 1)$ is measurable with respect to $\mathcal{F}_{i-1}$ and $B := \sup_{i\in\mathbb{N}} \frac{\alpha_i}{1-\lambda_i} < \infty$. The `online-adaptive` bound by Katsevich and Ramdas [18]

$$d^{\mathrm{ad}}(S_t) = \left\lceil -ca + \sum_{i=1}^{t} \mathbb{1}\{P_i \le \alpha_i\} - c\frac{\alpha_i}{1-\lambda_i}\mathbb{1}\{P_i > \lambda_i\} \right\rceil$$

provides simultaneous true discovery guarantee over all sets $S_t = \{i \le t : P_i \le \alpha_i\}$, $t \in \mathbb{N}$, where $a > 0$ is some regularization parameter and $c = \frac{\log(1/\alpha)}{a\log(1+(1-\alpha^{B/a})/B)}$. Note that c has a different value than for the `online-simple` algorithm. Similar as demonstrated in Section 3.2, Katsevich and Ramdas [18] proved the error guarantee by showing that $E_i^{\mathrm{ad}} = \exp[\theta_c(\mathbb{1}\{P_i \le \alpha_i\} - c\frac{\alpha_i}{1-\lambda_i}\mathbb{1}\{P_i > \lambda_i\})]$, $i \in \mathbb{N}$, are sequential e-values, where $\theta_c = \log(1/\alpha)/(ca)$. Note that

$$\prod_{i\in I} E_i^{\mathrm{ad}} \ge 1/\alpha \Leftrightarrow \sum_{i\in I} \mathbb{1}\{P_i \le \alpha_i\} - c\frac{\alpha_i}{1-\lambda_i}\mathbb{1}\{P_i > \lambda_i\} \ge \log(1/\alpha)/\theta_c = ca \quad (I \in 2^{\mathbb{N}_f}).$$

With this, one can define a uniform improvement of the `online-adaptive` algorithm in the exact same manner as for the `online-simple` algorithm. Note that the `online-adaptive` method already adapts to the proportion of null hypotheses using the parameter λ_i and therefore cannot be further improved by the (online) closure principle in that direction. However, it still leads to a real uniform improvement by transforming it into a coherent procedure.

Proposition S.1. *The `SeqE-Guard` algorithm applied with the sequential e-values $(E_i^{ad})_{i\in\mathbb{N}}$ uniformly improves the `online-adaptive` method by Katsevich and Ramdas [18].*

Furthermore, the e-values E_i^{ad} are inadmissible if $\alpha_i/(1-\lambda_i)$ is not constant for all $i \in \mathbb{N}$ and thus can be improved. For this, note that $E_i^{\mathrm{ad}} = \exp(\theta_c)$, if $P_i \le \alpha_i$, $E_i^{\mathrm{ad}} = 1$, if $\alpha_i < P_i \le \lambda_i$ and

$E_i^{\text{ad}} = \exp(-\theta_c c\alpha_i/(1-\lambda_i))$, if $P_i > \lambda_i$. Hence, for all $\mathbb{P} \in H_i$, we can provide a tight upper bound for the expectation of E_i^{ad} by

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[E_i^{\text{ad}}|\mathbb{F}_{i-1}] &= \exp(\theta_c)\mathbb{P}(P_i \le \alpha_i|\mathcal{F}_{i-1}) \\
&\quad + \mathbb{P}(\alpha_i < P_i \le \lambda_i|\mathcal{F}_{i-1}) + \exp\left(-\theta_c c\frac{\alpha_i}{1-\lambda_i}\right)\mathbb{P}(P_i > \lambda_i|\mathcal{F}_{i-1}) \\
&= (\exp(\theta_c)-1)\mathbb{P}(P_i \le \alpha_i|\mathcal{F}_{i-1}) + \mathbb{P}(P_i \le \lambda_i|\mathcal{F}_{i-1})\left[1-\exp\left(-\theta_c c\frac{\alpha_i}{1-\lambda_i}\right)\right] \\
&\quad + \exp\left(-\theta_c c\frac{\alpha_i}{1-\lambda_i}\right) \\
&\le (\exp(\theta_c)-1)\alpha_i + \lambda_i\left[1-\exp\left(-\theta_c c\frac{\alpha_i}{1-\lambda_i}\right)\right] + \exp\left(-\theta_c c\frac{\alpha_i}{1-\lambda_i}\right) \\
&=: u(\alpha, \alpha_i, \lambda_i, B, a),
\end{aligned}$$

which can easily be calculated for given α, α_i, λ_i, B and a. If $B = \alpha_i/(1-\lambda_i)$, then $u(\alpha, \alpha_i, \lambda_i, B, a) = 1$. However, in practice $\alpha_i/(1-\lambda_i)$ may vary over time such that there are indices $j \in \mathbb{N}$ with $\alpha_j/(1-\lambda_j) < B$. In this case, the e-value E_j^{ad} becomes conservative. For example, if $\alpha = 0.1$, $\alpha_i = 0.1$, $\lambda_i = 0.5$, $B = 0.4$ and $a = 1$, then $u(\alpha, \alpha_i, \lambda_i, B, a) = 0.966$.

S.3.2 Uniform improvements of the methods by Meah et al. [21]

We derive uniform improvements of the online procedures `u-online-simple` and `u-online-Freedman` of Meah et al. [21]. It should be noted that the uniform improvements presented here are not instances of our `SeqE-Guard` algorithm, however, they are obtained by taking the union of many `SeqE-Guard` bounds or by taking the mean of many test martingales (instead of a single one) for each intersection test, and hence are closely related to `SeqE-Guard`.

S.3.2.1 Uniform improvement of the `u-online-simple` method

Meah et al. [21] have introduced a modified version of the `online-simple` method by Katsevich and Ramdas [18]. In the following, we show how this method can be uniformly improved by e-value based online closed testing.

Let p-values $P_1, P_2, \ldots$ and significance levels $\alpha_1, \alpha_2, \ldots$ be defined as for the `online-simple` algorithm (see Section 3.2), and the null p-values be valid conditional on the past. Instead of fixing the parameter $a > 0$ in advance, Meah et al. [21] combined `online-simple` bounds for different parameters using a union bound. More precisely, for each $a \in \mathbb{N}$ let $\alpha(a) = \frac{6\alpha}{a^2\pi^2}$ and $c_a = \frac{\log(1/\alpha(a))}{a\log(1+\log(1/\alpha(a))/a)}$. By (13), we have that for all $\mathbb{P} \in \mathcal{P}$:

$$\mathbb{P}(\boldsymbol{d}_a^{\text{os}}(S_t) \le |S_t \cap I_1^{\mathbb{P}}| \text{ for all } t \in \mathbb{N}) \ge 1-\alpha(a), \tag{S.2}$$

$$\text{where } \boldsymbol{d}_a^{\text{os}}(S_t) = \left\lceil -c_a a + \sum_{i=1}^{t} \mathbb{1}\{P_i \le \alpha_i\} - c_a\alpha_i \right\rceil \quad (t \in \mathbb{N}), \tag{S.3}$$

with $\alpha_1, \alpha_2, \ldots$ being nonnegative thresholds and $S_t = \{i \le t : P_i \le \alpha_i\}$. Meah et al. [21] proposed

$$\boldsymbol{d}^{\text{u-os}}(S_t) = \max_{a\in\mathbb{N}} \boldsymbol{d}_a^{\text{os}}(S_t) \quad t \in \mathbb{N}, \tag{S.4}$$

whose true discovery guarantee follows by applying a union bound to (S.2). We refer to the procedure $\boldsymbol{d}^{\text{u-os}}$ as `u-online-simple` method in this paper.

In Section 3.2 we have shown that the bound in (S.3) can be uniformly improved by Algorithm 2 for all $a > 0$ and $\alpha(a) \in (0, 1)$. Hence, a simple uniform improvement of the `u-online-simple` method can be obtained by taking the maximum of these improved bounds.

More precisely, let $E_i^{\text{os},a} = \exp[\theta_{c_a}(\mathbb{1}\{P_i \le \alpha_i\} - c_a\alpha_i)]$ and define $W_I^{t,a} := \prod_{i \in I \cap \{1,\dots,t\}} E_i^{\text{os},a}$. Then the `closed u-online-simple` method is given by $\boldsymbol{d}^{\text{cu-os}}(S_t) = \max_{a\in\mathbb{N}} \boldsymbol{d}^{\boldsymbol{\phi}^a}(S_t)$, where $\boldsymbol{d}^{\boldsymbol{\phi}^a}$ is the online closed procedure obtained by the intersection tests

$$\phi_I^a := \mathbb{1}\left\{\exists t \in I : W_I^{t,a} \ge 1/\alpha(a)\right\} = \mathbb{1}\left\{\exists t \in I : \frac{6}{a^2\pi^2} W_I^{t,a} \ge 1/\alpha\right\}, \quad a \in \mathbb{N}, I \subseteq \mathbb{N}.$$

Since $\boldsymbol{d}^{\boldsymbol{\phi}^a}(S_t) \ge \boldsymbol{d}_a^{\text{os}}(S_t)$ for all $a \in \mathbb{N}$ (see Section 3.2), we also have $\boldsymbol{d}^{\text{cu-os}}(S_t) \ge \boldsymbol{d}^{\text{u-os}}(S_t)$.

However, we can improve $\boldsymbol{d}^{\text{cu-os}}$ even further. For this, note that ϕ_I^a can be uniformly improved by

$$\phi_I^{\text{m-os}} := \mathbb{1}\left\{\exists t \in I : \sum_{a\in\mathbb{N}} \frac{6}{a^2\pi^2} W_I^{t,a} \ge 1/\alpha\right\},$$

simultaneously for all a. Since (weighted) means of test supermartingales are test supermartingales again, Ville's inequality implies that $\phi_I^{\text{m-os}}$ is an intersection test and the online closed procedure $\boldsymbol{d}^{\text{m-os}} := \boldsymbol{d}^{\boldsymbol{\phi}^{\text{m-os}}}$ provides simultaneous true discovery guarantee. We refer to $\boldsymbol{d}^{\text{m-os}}$ as `m-online-simple` procedure in the following. Since $\phi_I^{\text{m-os}} \ge \phi_I^a$ for all $a \in \mathbb{N}$, we have

$$\boldsymbol{d}^{\text{m-os}}(S_t) \ge \boldsymbol{d}^{\text{cu-os}}(S_t) \ge \boldsymbol{d}^{\text{u-os}}(S_t) \text{ for all } t \in \mathbb{N}.$$

Note that $E_i^{\text{os},a'} \ge E_j^{\text{os},a'}$ for some $a' \in \mathbb{N}$ and $i \ne j$ implies that $E_i^{\text{os},a} \ge E_j^{\text{os},a}$ for all $a \in \mathbb{N}$ and $E_i^{\text{os},a'} \ge 1$ for some $a' \in \mathbb{N}$ and $i \in \mathbb{N}$ implies that $E_i^{\text{os},a} \ge 1$ for all $a \in \mathbb{N}$. With this, we can derive a short-cut (see Algorithm 3) for $\boldsymbol{d}^{\text{m-os}}(S_t)$, $t \in \mathbb{N}$, in the same manner as we did in Algorithms 1 and 2.

Proposition S.2. *The `m-online-simple` procedure (Algorithm 3) and the `cu-online-simple` procedure uniformly improve the `u-online-simple` method by Meah et al. [21].*

In Figure S.1, we compare the `u-online-simple` method [21] with our `m-online-simple` procedure (Algorithm 3). The simulation setup is the same as described in Section 4. The plots show that the improvement obtained by the `m-online-simple` procedure is substantial.

S.3.2.2 Uniform improvement of the `u-online-Freedman` method

Similarly as the `u-online-simple` method (Section S.3.2.1), the `u-online-Freedman` procedure is obtained by taking the (weighted) union of many individual bounds. However, instead of using the `online-simple` bounds by Katsevich and Ramdas [18], the `u-online-Freedman` procedure is based on Freedman's inequality [7].

Let p-values $P_1, P_2, \dots$ and significance levels $\alpha_1, \alpha_2, \dots$ be defined as for the `online-simple` algorithm (see Section 3.2), and the null p-values be valid conditional on the past. Meah et al. [21] showed (see Corollary 41 in [21]) that for all $\mathbb{P} \in \mathcal{P}$:

$$\mathbb{P}(\boldsymbol{d}_a^{\text{Freed}}(S_t) > |S_t \cap I_1^{\mathbb{P}}| \text{ and } A_t \le a \text{ for some } t \in \mathbb{N}) \le \alpha(a), \tag{S.5}$$

$$\text{where } \boldsymbol{d}_a^{\text{Freed}}(S_t) = 1 + \left\lfloor -\kappa_a + \sum_{i=1}^{t} \mathbb{1}\{P_i \le \alpha_i\} - \alpha_i \right\rfloor, \tag{S.6}$$

$$\text{and } B_t = \sum_{i=1}^{t} \alpha_i(1 - \alpha_i), \tag{S.7}$$

Algorithm 3 `m-online-simple`

Input: Sequence of p-values $P_1, P_2, \ldots$ and sequence of (potentially data-dependent) individual significance levels $\alpha_1, \alpha_2, \ldots$.

Output: Query sets $S_1 \subseteq S_2 \subseteq \ldots$ and true discovery bounds $d_1 \leq d_2 \leq \ldots$.

$d_0 = 0$
$S_0 = \emptyset$
$A_0^c = \emptyset$
for $t = 1, 2, \ldots$ **do**
  if $P_t \leq \alpha_t$ **then**
    $S_t = S_{t-1} \cup \{t\}$
  else
    $S_t = S_{t-1}$
  end if
  if $\sum_{a \in \mathbb{N}} \frac{6}{a^2\pi^2} \prod_{i \in \{1,\ldots,t\} \setminus A_{t-1}^c} E_i^{\mathrm{os},a} \geq 1/\alpha$ **then**
    $d_t = d_{t-1} + 1$
    $A_t^c = A_{t-1}^c \cup \{\text{index of smallest individual significance level in } S_t \setminus A_{t-1}^c\}$
  else
    $d_t = d_{t-1}$
    $A_t^c = A_{t-1}^c$
  end if
  return S_t, d_t
end for

with $S_t = \{i \leq t : P_i \leq \alpha_i\}$, $\kappa_a = \sqrt{2a\log(1/\alpha(a))} + \frac{\log(1/\alpha(a))}{2}$ and $\alpha(a) = \alpha\left(\frac{6}{\max(2\log_2(a),1)^2(\pi^2+6)}\right)$ for some parameter $a = 2^{j/2}, j \in \mathbb{N} \cup \{0\}$. The `u-online-Freedman` procedure is then obtained by applying a union to (S.5) over all $j \in \mathbb{N} \cup \{0\}$ (see Corollary 42 in [21]). In the following, we show that the `SeqE-Guard` algorithm allows to uniformly improve the bound $\boldsymbol{d}_a^{\mathrm{Freed}}$ for each a.

Howard et al. [15] improved Freedman's inequality by a test martingale approach. Our improvement of the bound $\boldsymbol{d}_a^{\mathrm{Freed}}$ is based on the same technique, which additionally uses the `SeqE-Guard` algorithm. For this, we define the sequential e-values

$$E_i^{\mathrm{Freed},a} \coloneqq \exp(\lambda_a(\mathbb{1}\{P_i \leq \alpha_i\} - \alpha_i) - \psi(\lambda_a)\alpha_i(1-\alpha_i)) \quad (i \in \mathbb{N}), \tag{S.8}$$

where $\lambda_a = \log\left(1 + \frac{\kappa_a}{a}\right)$ and $\psi(\lambda_a) = \exp(\lambda_a) - \lambda_a - 1$. To see that $E_i^{\mathrm{Freed},a}$, $i \in \mathbb{N}$, are sequential e-values, define $X_i \coloneqq \mathbb{1}\{P_i \leq \alpha_i\} - \alpha_i$ and note that for all $i \in I_0^{\mathbb{P}}$ it holds that $X_i \leq 1$, $\mathbb{E}_{\mathbb{P}}[X_i|\mathcal{F}_{i-1}] \leq 0$ and $\mathbb{E}_{\mathbb{P}}[X_i^2|\mathcal{F}_{i-1}] \leq \alpha_i(1-\alpha_i)$. Therefore, $\mathbb{E}_{\mathbb{P}}[\exp(\lambda_a X_i)|\mathcal{F}_{i-1}] \leq \exp(\psi(\lambda_a)\alpha_i(1-\alpha_i))$ (e.g., Lemma 6.7 in [33]). Now let $d_t^{\mathrm{Freed},a}$ be the bound obtained by applying `SeqE-Guard` with $E_1^{\mathrm{Freed},a}, E_2^{\mathrm{Freed},a}, \ldots$. In the same manner as in (15), we get that

$$\begin{aligned} d_t^{\mathrm{Freed},a} &\geq 1 + \left\lfloor -\frac{\log(1/\alpha(a))}{\lambda_a} + \sum_{i=1}^{t} \mathbb{1}\{P_i \leq \alpha_i\} - \alpha_i - \frac{\psi(\lambda_a)\alpha_i(1-\alpha_i)}{\lambda_a} \right\rfloor \\ &= 1 + \left\lfloor -\frac{\log(1/\alpha(a)) + \psi(\lambda_a)B_t}{\lambda_a} + \sum_{i=1}^{t} \mathbb{1}\{P_i \leq \alpha_i\} - \alpha_i \right\rfloor . \end{aligned}$$

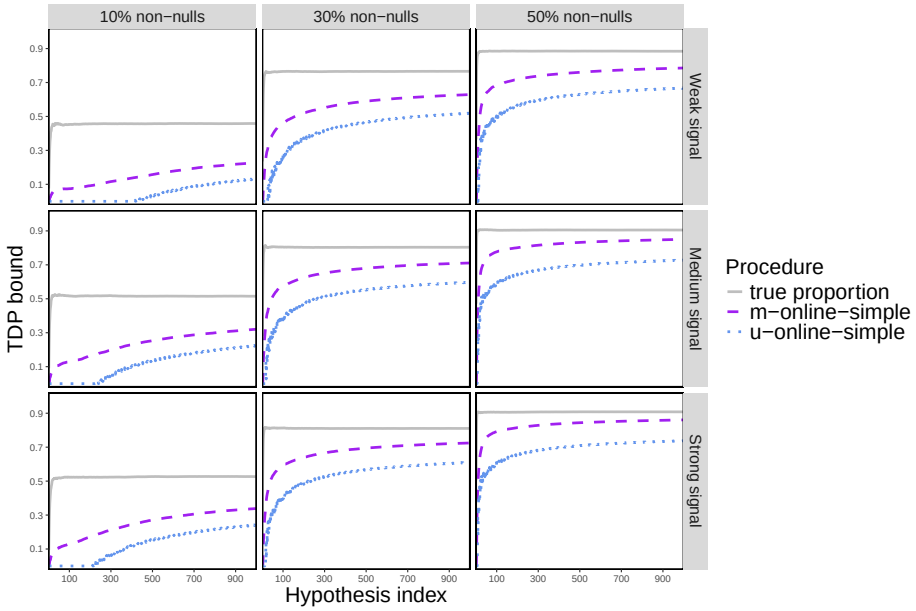


Figure S.1: True discovery proportion bounds obtained by the `u-online-simple` method [21] and our `m-online-simple` method. The `m-online-simple` procedure improves the `u-online-simple` method substantially.

Further, if $B_t \leq a$, we obtain

$$\begin{aligned}
&- \frac{\log(1/\alpha(a)) + \psi(\lambda_a) B_t}{\lambda_a} \\
&\geq - \frac{\log(1/\alpha(a)) + \psi(\lambda_a) a}{\lambda_a} \\
&= - \frac{\log(1/\alpha(a)) + \kappa_a - \log\left(1 + \frac{\kappa_a}{a}\right)(a + \kappa_a) + \log\left(1 + \frac{\kappa_a}{a}\right)\kappa_a}{\log\left(1 + \frac{\kappa_a}{a}\right)} \\
&\geq - \frac{\log(1/\alpha(a)) - 2(\sqrt{a + \kappa_a} - \sqrt{a})^2 + \log\left(1 + \frac{\kappa_a}{a}\right)\kappa_a}{\log\left(1 + \frac{\kappa_a}{a}\right)},
\end{aligned}$$

where the second inequality follows by Lemma 43 of Meah et al. [21]. This can be further simplified by

$$
\begin{aligned}
&- \frac{\log(1/\alpha(a)) - 2(\sqrt{a+\kappa_a} - \sqrt{a})^2 + \log\left(1 + \frac{\kappa_a}{a}\right)\kappa_a}{\log\left(1 + \frac{\kappa_a}{a}\right)} \\
&= -\frac{\log(1/\alpha(a)) - 2(\sqrt{a+\kappa_a} - \sqrt{a})^2}{\log\left(1 + \frac{\kappa_a}{a}\right)} - \kappa_a \\
&= -\frac{\log(1/\alpha(a)) - 2\left(\sqrt{a + \sqrt{2a\log(1/\alpha(a))} + \frac{\log(1/\alpha(a))}{2}} - \sqrt{a}\right)^2}{\log\left(1 + \frac{\kappa_a}{a}\right)} - \kappa_a \\
&= -\frac{\log(1/\alpha(a)) - 2\left(\sqrt{\left(\sqrt{a} + \sqrt{\frac{\log(1/\alpha(a))}{2}}\right)^2} - \sqrt{a}\right)^2}{\log\left(1 + \frac{\kappa_a}{a}\right)} - \kappa_a \\
&= -\frac{\log(1/\alpha(a)) - 2\left(\sqrt{\frac{\log(1/\alpha(a))}{2}}\right)^2}{\log\left(1 + \frac{\kappa_a}{a}\right)} - \kappa_a \\
&= -\kappa_a,
\end{aligned}
$$

which shows that $d_t^{\text{Freed},a} \geq \boldsymbol{d}_a^{\text{Freed}}(S_t)$ for all $a \geq B_t$ and we additionally improve (S.5) by providing a nontrivial bound for $a < B_t$. Hence, a uniform improvement of the `u-online-Freedman` procedure can either be obtained by taking the maximum of the improved bounds $d_t^{\text{Freed},a}$ over all $a = 2^j$, $j \in \mathbb{N} \cup \{0\}$, or by the more powerful mean strategy described in Supplementary Material S.3.2.1. In line with Supplementary Material S.3.2.1, we call these two procedures `cu-online-Freedman` and `m-online-Freedman`, respectively.

Proposition S.3. *The `m-online-Freedman` procedure and the `cu-online-Freedman` procedure uniformly improve the `u-online-Freedman` method by Meah et al. [21].*

Note that the sequential e-values $E_i^{\text{Freed},a}$, $i \in \mathbb{N}$, are just binary e-values depending on $\mathbb{1}\{P_i \leq \alpha_i\}$. Hence, they are very similar to the ones defined for the `online-simple` method (14), the difference lies just in the weighting of the two cases $P_i \leq \alpha_i$ and $P_i > \alpha_i$. Therefore, the `SeqE-Guard` algorithm does not only improve these methods, it also facilitates their interpretation. We only need to ensure that the expected value of each sequential e-value is bounded by one under the null hypothesis, which is very easy to check. Indeed, one can check that $E_i^{\text{Freed},a}$ is slightly conservative as well and could be improved, we just chose the representation (S.8) as it simplifies the proof of the uniform improvement. Such looseness would be difficult to detect with the procedures by Katsevich and Ramdas [18] and Meah et al. [21], as their proofs of validity are based on far more complicated arguments.

In Figure S.2, we compare the `u-online-Freed` method [21] with our `m-online-Freed` procedure. The simulation setup is the same as described in Section 4. The behavior of the `u-online-Freed` and `m-online-Freed` method is similar as for the `u-online-simple` and `m-online-simple` method (see Figure S.1). The `m-online-Freed` method leads to a significant uniform improvement over the `u-online-Freed` method.

S.4 Calibrating sequential p-values into e-values

While e-values have been of recent interest, most studies still use p-values as measure of evidence against the null hypothesis. In this case, we can *calibrate* the p-values into e-values and then apply `SeqE-Guard`. In this context, a calibrator is a function from p-values to e-values, meaning it takes as an input a p-value and yields an e-value as an output.

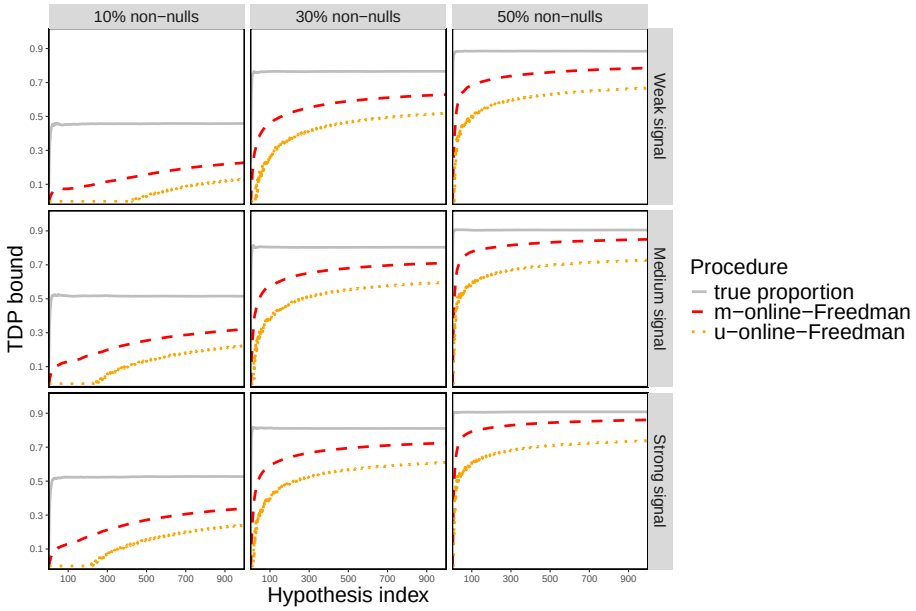


Figure S.2: True discovery proportion bounds obtained by the `u-online-Freed` method [21] and our `m-online-Freed` method. The `m-online-Freed` procedure improves the `u-online-Freed` method substantially.

For example, the `online-simple` method of Katsevich and Ramdas [18] introduced in Section 3.2 is implicitly based on calibrating each p-value into a simple binary e-value. However, there are infinitely many other calibrators that could be used. Assume that p-values $(P_t)_{t\in\mathbb{N}}$ for the hypotheses $(H_t)_{t\in\mathbb{N}}$ are given. A decreasing function $f : [0,1] \to [0,\infty]$ is a calibrator if $\int_0^1 f(x)dx \leq 1$, and it is admissible if equality holds [39, 36]. Note that if the p-values are sequential p-values, meaning P_t is measurable with respect to $\mathcal{F}_t$ and $\mathbb{P}(P_t \leq x|\mathcal{F}_{t-1}) \leq x$ for all $x \in [0,1]$, then the calibrated e-values are sequential e-values.

An example calibrator is

$$h_x(p) = \exp\left(x\Phi^{-1}(1-p) - x^2/2\right), \tag{S.9}$$

where Φ denotes the CDF of a standard normal distribution. To see that this is a valid calibrator, note that $\Phi^{-1}(U)$, where $U \sim [0,1]$, follows a standard normal distribution and the moment generating function of a standard normal distribution for the real parameter $x > 0$ is given by $\exp(x^2/2)$. Duan et al. [4] used this calibrator to derive a martingale Stouffer [32] global test. This global test is based on a confidence sequence of Howard et al. [15].

We compare the application of `SeqE-Guard` with calibrated e-values and hedged GRO e-values experimentally in the next section.

S.5 Boosting of sequential e-values

Wang and Ramdas [41] introduced a way to *boost* e-values before plugging them into their e-BH procedure without violating the desired FDR control. In this section, we propose a similar (but simpler) approach for `SeqE-Guard` that will improve its power for a specific query path.

We begin by noting that whenever the bound d_t is increased by one, the largest e-value with index in A will not be considered in the following analysis. Hence, extremely large e-values will be

excluded from the analysis anyway, which makes it possible to truncate the e-values at a specific threshold without changing its outcome. This makes the resulting truncated e-values conservative under the null, and one can improve the procedure by multiplying the e-value by a suitable constant larger than 1 to remove its conservativeness. This truncation+multiplication operation is what is referred to as *boosting* the e-value.

To this end, recall that $A_{t-1} \subseteq S_{t-1}$, $t \in \mathbb{N}$, is the index set of all previous e-values in the query set that were not already excluded by `SeqE-Guard` and $U_{t-1} \subseteq \{1, \dots, t-1\} \setminus S_{t-1}$ the index set of all previous e-values that are not contained in the query set and that are smaller than 1. Now, define

$$m_t := \max \left\{ \max_{i \in A_{t-1}} E_i, \frac{1}{\alpha \prod_{i \in A_{t-1} \cup U_{t-1}} E_i} \right\}, \tag{S.10}$$

and note that m_t is predictable (measurable with respect to $\mathcal{F}_{t-1}$). Furthermore, if $E_t \geq m_t$ and $t \in S_t$, then $d_t = d_{t-1} + 1$ and E_t will be excluded in the further analysis. If $E_t \geq m_t$ and $t \notin S_t$, then $E_t \geq 1$, since $m_t \geq 1$ by definition, and E_t won't be considered in the analysis anyway. Hence, we define the truncation function $T_t : [0, \infty] \to [0, m_t]$ as

$$T_t(x) := x \mathbb{1}\{x \leq m_t\} + m_t \mathbb{1}\{x > m_t\} \tag{S.11}$$

and then choose a boosting factor $b_t \geq 1$ as large as possible such that

$$\mathbb{E}_{\mathbb{P}}[T_t(b_t E_t) | \mathcal{F}_{t-1}] \leq 1 \text{ for all } \mathbb{P} \in H_t. \tag{S.12}$$

Note that $b_t = 1$ always satisfies (S.12); so a boosting factor always exists and is always at least one. Condition (S.12) immediately implies that $T_t(b_t E_t)$ is a sequential e-value. Furthermore, using $b_t E_t$ in `SeqE-Guard` yields exactly the same results as using $T_t(b_t E_t)$. Therefore, applying `SeqE-Guard` to the boosted e-values $(b_t E_t)_{t \in \mathbb{N}}$ provides simultaneous true discovery guarantee and is uniformly more powerful than with non-boosted e-values, since $b_t \geq 1$. Note that in this case m_t should also be calculated based on the boosted e-values $b_1 E_1, \dots, b_{t-1} E_{t-1}$. As also mentioned by Wang and Ramdas [41], one could use different functions than $x \mapsto bx$ for some $b \geq 1$ to boost the e-values. In general, it is only required that each boosted e-value E_t^{boost}, $t \in \mathbb{N}$, satisfies $\mathbb{E}_{\mathbb{P}}[T_t(E_t^{\text{boost}}) | \mathcal{F}_{t-1}] \leq 1$ for all $\mathbb{P} \in H_t$. We summarize this result in the following theorem.

Proposition S.1. *Let $E_1^{\text{boost}}, E_2^{\text{boost}}, \dots$ be a sequence of nonnegative random variables such that $\mathbb{E}_{\mathbb{P}}[T_t(E_t^{\text{boost}}) | \mathcal{F}_{t-1}] \leq 1$ for all $\mathbb{P} \in H_t$, where T_t is given by* (S.11) *and m_t is calculated as in* (S.10) *based on $E_1^{\text{boost}}, \dots, E_{t-1}^{\text{boost}}$. Then, applying `SeqE-Guard` to the boosted sequential e-values $E_1^{\text{boost}}, E_2^{\text{boost}}, \dots$ provides simultaneous true discovery guarantee.*

It should be noted that boosting adapts the sequential e-values to the query path $(S_t)_{t \geq 1}$ and Proposition S.1 only holds for the query path for which the boosted e-values $E_1^{\text{boost}}, E_2^{\text{boost}}, \dots$ were constructed. If multiple query paths are investigated, one must use the sequential e-values $T_1(E_1^{\text{boost}}), T_2(E_2^{\text{boost}}), \dots$ for all other query paths. Therefore, it may happen that boosting weakens the performance of `SeqE-Guard` for the other query paths.

In the following we provide several examples that illustrate how the boosting factors can be determined in specific cases and demonstrate the possible gain in efficiency.

Example S.1. *We consider Example 3 from Wang and Ramdas [41] adapted to our setting. For each $t \in \mathbb{N}$, we test the simple null hypothesis $H_t : X_t | \mathcal{F}_{t-1} \sim \mathcal{N}(\mu_0, 1)$ against the simple alternative $H_t^A : X_t | \mathcal{F}_{t-1} \sim \mathcal{N}(\mu_1, 1)$, where X_t denotes the data for H_t. In this case, the GRO e-value is given by the likelihood ratio between two normal distributions with variance 1 and means μ_1 and μ_0*

$$E_t = \exp(\delta Z_t - \delta^2/2), \tag{S.13}$$

where $\delta = \mu_1 - \mu_0 > 0$ and $Z_t = X_t - \mu_0$ follows a standard normal distribution conditional on $\mathcal{F}_{t-1}$ under H_t. Hence, conditional on the past, each null e-value follows a log-normal distribution with parameter $(-\delta^2/2, \delta)$. With this, we obtain for all $t \in I_0^{\mathbb{P}}$:

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}[b_t E_t \mathbb{1}\{b_t E_t \leq m_t\} + m_t \mathbb{1}\{b_t E_t > m_t\} | \mathcal{F}_{t-1}] \\
&= b_t \mathbb{E}_{\mathbb{P}}[E_t \mathbb{1}\{b_t E_t \leq m_t\} | \mathcal{F}_{t-1}] + m_t \mathbb{P}(b_t E_t > m_t | \mathcal{F}_{t-1}) \\
&= b_t \left[1 - \Phi\left(\frac{\delta}{2} - \frac{\log\left(m_t/b_t\right)}{\delta}\right)\right] + m_t \left[1 - \Phi\left(\frac{\log(m_t/b_t) + \delta^2/2}{\delta}\right)\right],
\end{aligned}$$

where Φ is the CDF of a standard normal distribution. The last expression can be set equal to 1 and then be solved for b_t numerically. For example, for $\delta = 3$ and $m_t = 20$, we obtain $b_t = 3.494$. Hence, the e-value E_t could be multiplied by 3.494 without violating the true discovery guarantee, a substantial gain. In general, the larger m_t, the smaller is the boosting factor. For example, if $m_t = 5$, then $b_t = 11.826$ and if $m_t = 100$, then $b_t = 1.774$. Nevertheless, even the latter boosting factor would increase the power of the true discovery procedure significantly and we would usually expect m_t to be smaller than 100 in most settings. If we use, as described in Section 3.3, the e-value $1 - \lambda_t + \lambda_t E_t^{\text{GRO}}$, $\lambda_t \in (0, 1)$ instead, we need to solve

$$m_t + \Phi\left(\frac{\log(s_t) + \delta^2/2}{\delta}\right)[b_t(1 - \lambda_t) - m_t] + b_t \lambda_t \left[1 - \Phi\left(\frac{\delta}{2} - \frac{\log\left(s_t\right)}{\delta}\right)\right] = 1,$$

for $b_t \in [1, 1/(1 - \lambda_t))$, where $s_t = (\lambda_t - 1 + m_t/b_t)/\lambda_t$. In this case, $\delta = 3$, $\lambda_t = 0.5$ and $m_t = 20$ yield a boosting factor of $b_t = 1.354$.

Example S.2. *Suppose we observe sequential p-values $P_1, P_2, \ldots$ and want to apply the calibrator* (S.9)*. If the p-values are uniformly distributed conditional on the past, the resulting e-value has the exact same distribution as the e-value in* (S.13) *under the null hypothesis for $\delta = x$, where x is the freely chosen parameter for the calibrator. Hence, we can do the exact same calculations to obtain an appropriate boosting factor. If the p-values are stochastically larger than uniform, we could still use that same boosting factor, as the resulting e-values provide true discovery guarantee but might be conservative.*

Example S.3. *In case of the `closed online-simple` method (Algorithm 2) it is particular simple to "boost" the e-values. Since E_i^{os} only takes two different values, we can simply ensure $E_t^{\text{os}} \leq m_t$ by choosing α_t such that $E_t^{\text{os}} \leq m_t$ if $\mathbb{1}\{P_t \leq \alpha_t\} = 1$. Note that in case of $\alpha_t = \nu$ for all $t \in \mathbb{N}$ and some $\nu > 0$ such that $\exp[\theta_c(1 - c\nu)] \leq 1/\alpha$ it is not possible to improve the bounds of the `closed online-simple` method further by boosting, since we already have $E_t^{\text{os}} \leq m_t$ almost surely.*

In Figure S.3 we compare the `admissible online-simple` method with the `SeqE-Guard` algorithm with boosted GRO (the GRO e-values were first hedged and then boosted as described in Example S.1) and calibrated e-values (using the calibrator in (S.9) with $x = 0.1$ and boosting as described in Example S.2). The calibrated e-values perform worst in most cases, boosted GRO e-values perform best if the proportion of false hypotheses is small and the `admissible online-simple method` performs best if the proportion of false hypotheses is large. Also note that compared to Figure 6, the bound of the hedged GRO e-values increased substantially using boosting. However, it adapts the e-values to the query paths and thus the bounds might decrease for other query paths.

S.6 Analysis of IMPC data

The International Mouse Phenotyping Consortium (IMPC) [22] is a large-scale effort to systematically characterize the phenotypic consequences of knocking out each protein-coding gene in the mouse genome. Since new genes are analyzed and released over time, the resulting sequential structure

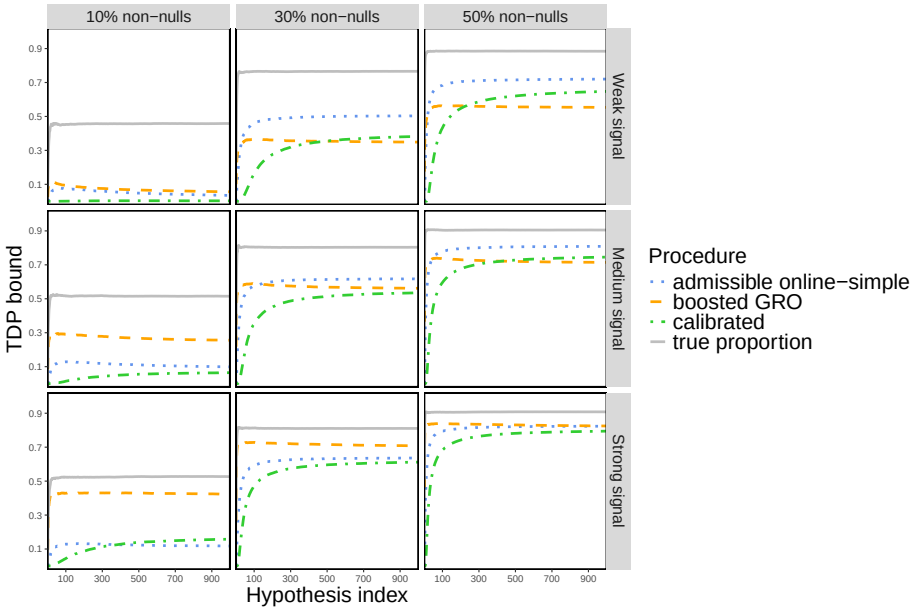


Figure S.3: True discovery proportion bounds obtained by the `admissible online-simple` as well as the `SeqE-Guard` algorithm with boosted GRO and (boosted) calibrated e-values. The boosted GRO e-values perform best if the proportion of false hypotheses is small and the `admissible online-simple` method performs best if the proportion is large.

naturally gives rise to an online multiple testing problem, making the IMPC dataset a standard application for evaluating online procedures [28]. In this section, we apply the proposed methodology to 172 081 p-values that resulted from an evaluation by Karp et al. [17] and that are available at the Zenodo repository `zenodo.org/records/2396572` [27].

In our analysis, we compare the `online-simple` procedure by Katsevich and Ramdas [18] with our `admissible online-simple` method when applied to the query path obtained by the LORD++ algorithm [16, 24]. The LORD++ algorithm is an online procedure with FDR control that produces the rejection set $S_t = \{i \leq t : P_i \leq \alpha_i\}$, where the nonnegative thresholds $\alpha_1, \alpha_2, \ldots$ are chosen such that

$$\widehat{\mathrm{FDP}}_t := \frac{\sum_{i=1}^{t} \alpha_i}{\max(|S_t|, 1)} \leq \alpha_{\mathrm{FDR}}.$$

The idea is to conservatively estimate $\mathrm{FDP}(S_t)$ by $\widehat{\mathrm{FDP}}_t$ and then choose the levels $\alpha_1, \alpha_2, \ldots$ such that $\sup_{t \in \mathbb{N}} \widehat{\mathrm{FDP}}_t \leq \alpha_{\mathrm{FDR}}$, where α_{FDR} is the predefined nominal FDR level [24]. Interestingly, the bound of the `online-simple` method was constructed using the estimate $\widehat{\mathrm{FDP}}_t$ for $\mathrm{FDP}(S_t)$ [18]. The LORD++ algorithm controls the FDR at level α_{FDR} if the p-values are independent. Please note that this assumption is not necessarily fulfilled for the IMPC dataset and the following analysis is for illustrative purposes only. The independence assumption also ensures that the p-values are valid conditional on the past and thus that the `online-simple` procedure and the `admissible online-simple` method provide simultaneous true discovery guarantee.

The results of our analysis are illustrated in Figure S.4. The FDP bounds were simply obtained by $\boldsymbol{q}(S_t) = 1 - \frac{\boldsymbol{d}(S_t)}{|S_t|}$, where $\boldsymbol{d}$ is the lower bound for the number of true discoveries. Both online true discovery procedures were applied with the same parameters as in Section 4 and the LORD++ algorithm with the standard values from the onlineFDR `R` package [27] but $\gamma_i = \frac{6}{\pi^2 i^2}$. Note that the level α for the true discovery procedures is to be distinguished from the FDR level α_{FDR} used

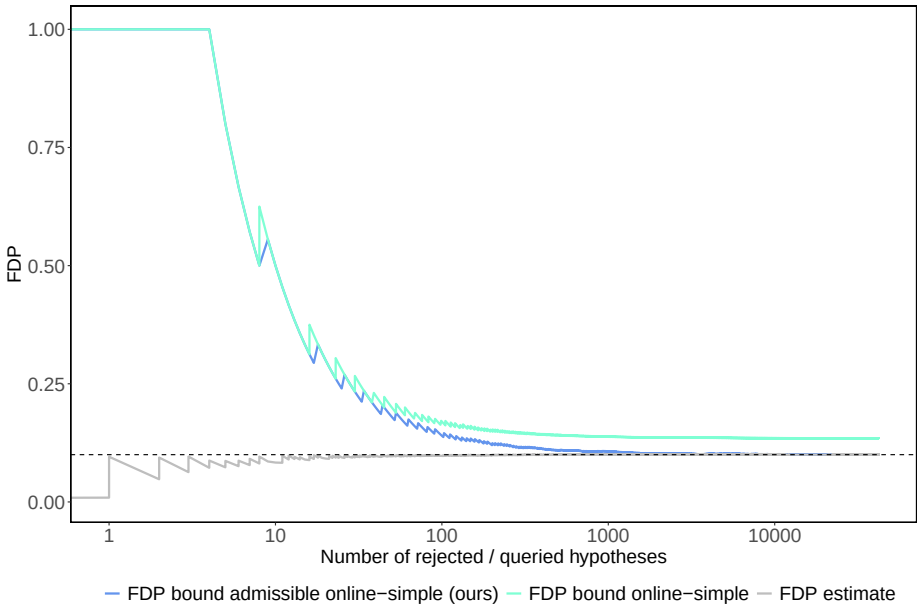


Figure S.4: FDP bounds obtained by the `online-simple` method [18] and our `admissible online-simple` method applied on IMPC data. The `admissible online-simple` improves the `online-simple` method substantially and is close to the nominal level FDR α_{FDR} of the LORD++ procedure.

to obtain the query path $(S_t)_{t\in\mathbb{N}}$ with the LORD++ algorithm. However, in this analysis both parameters were set to 0.1.

The results show that the `admissible online-simple` method substantially improves the online-simple procedure by yielding lower FDP bounds. Furthermore, the claims made by the `admissible online-simple` procedure are more informative than those made by the LORD++ algorithm for two reasons.

1. Since $\boldsymbol{q}(S_t) \approx \alpha_{\text{FDR}} = 0.1$ for $|S_t| \geq 1000$ for the `admissible online-simple` procedure $\boldsymbol{q}$, we can claim that the probability of the $\text{FDP}(S_t)$ being below 0.1 is greater than or equal to 0.9. In contrast, the LORD++ algorithm only allows to conclude that the expectation of $\text{FDP}(S_t)$ is bounded by 0.1. Even though the former statement does not imply the latter (and vice versa), the high probability statement often gives stronger confidence about the true $\text{FDP}(S_t)$ than the expectation bound.

2. The probability guarantee of the `admissible online-simple` method holds simultaneously for all t. In particular, this allows to stop the testing process data-adaptively while the claim made in the previous point remains valid. In contrast, the LORD++ algorithm is only proven to control the FDR at fixed times t, and thus the FDR could be inflated at data-adaptive stopping times.

S.7 Proof of Theorem 3.1

Proof. Let d_t, $t \in \mathbb{N}$, be the bounds of `SeqE-Guard`, Π_Z, $Z \subseteq \mathbb{N}$, be the product of all e-values with index in Z, $A_t \subseteq S_t$ (resp. $A_{t+} \subseteq S_t$) be the set A at step t before (resp. after) checking whether $\Pi_{i\in A\cup U} E_i \geq 1/\alpha$, $U_t \subseteq \{1,\dots,t\} \setminus S_t$ be the index set of e-values that are smaller than 1.

Recall, that $\boldsymbol{d}^{\boldsymbol{\phi}}(S)$ is solely determined by ϕ_I with $I \subseteq \{1, \dots, \max(S)\}$, since $\boldsymbol{\phi}$ is increasing (see (6)). Hence, we particularly have $d_1 = \boldsymbol{d}^{\boldsymbol{\phi}}(S_1)$. Now assume that $d_1 = \boldsymbol{d}^{\boldsymbol{\phi}}(S_1), \dots, d_{t-1} = \boldsymbol{d}^{\boldsymbol{\phi}}(S_{t-1})$ and $S_t = S_{t-1} \cup \{t\}$. Due to the coherence of $\boldsymbol{d}^{\boldsymbol{\phi}}$, it holds $d_{t-1} \leq \boldsymbol{d}^{\boldsymbol{\phi}}(S_t) \leq d_{t-1} + 1$. In the following, we show that $\Pi_{A_t \cup U_t} \geq 1/\alpha$ implies that $\boldsymbol{d}^{\boldsymbol{\phi}}(S_t) \geq d_{t-1} + 1$ and $\Pi_{A_t \cup U_t} < 1/\alpha$ implies that $\boldsymbol{d}^{\boldsymbol{\phi}}(S_t) < d_{t-1} + 1$, which proves the assertion.

We first show that $\boldsymbol{d}^{\boldsymbol{\phi}}(S_t) \geq d_{t-1} + 1$, if $\Pi_{A_t \cup U_t} \geq 1/\alpha$. For this, we prove that $\Pi_{A_t \cup U_t} \geq 1/\alpha$ implies $\phi_I = 1$ for all $I = V \cup W$, where $V \subseteq S_t$ and $W \subseteq \{1, \dots, t\} \setminus S_t$ with $|V| \geq |A_t|$. Since $\boldsymbol{\phi}$ is increasing, and thus $\boldsymbol{d}^{\boldsymbol{\phi}}(S_t)$ can be written by (6), this then implies that $\boldsymbol{d}^{\boldsymbol{\phi}}(S_t) \geq |S_t| - |A_t| + 1 = d_{t-1} + 1$. First, note that it is sufficient to show the claim for all $I = V \cup U_t$ with $|V| \geq |A_t|$, since multiplication with e-values that are larger than or equal to 1 cannot decrease the product. Now let such an I and V be fixed. Because of definition (12) it is necessary and sufficient to find some $t' \leq t$ such that $\Pi_{I \cap \{1,\dots,t'\}} \geq 1/\alpha$ (it will be shown that $t' = t_{\tilde{m}}$ does the trick, where $t_{\tilde{m}}$ is defined below).

Let $t_1, \dots, t_m$, where $t_m = t_{|S_t|-|A_t|+1} = t$, be the times at which $\Pi_{A_{t_i} \cup U_{t_i}} \geq 1/\alpha$ and $\tilde{m} \in \{1, \dots, m\}$ be the smallest index such that $|V \cap \{1, \dots, t_{\tilde{m}}\}| > |S_{t_{\tilde{m}}}| - \tilde{m}$. Note that $\tilde{m}$ always exists, because $|A_t| = |A_{t_m}| = |S_{t_m}| - m + 1$. With this, we have

$$\Pi_{I \cap \{1,\dots,t_{\tilde{m}}\}} \overset{(a)}{=} \Pi_{(V \cap \{1,\dots,t_{\tilde{m}}\}) \cup U_{t_{\tilde{m}}}} \overset{(b)}{\geq} \Pi_{(A_{t_{\tilde{m}}} \cap \{1,\dots,t_{\tilde{m}-1}\}) \cup (\{t_{\tilde{m}-1}+1,\dots t_{\tilde{m}}\} \cap S_t) \cup U_{t_{\tilde{m}}}} \overset{(c)}{=} \Pi_{A_{t_{\tilde{m}}} \cup U_{t_{\tilde{m}}}} \overset{(d)}{\geq} 1/\alpha.$$

As suggested by a referee, the individual steps of this derivation can be explained as follows:

(a) With $I = V \cup U_t$, it holds that $I \cap \{1, \dots, t_{\tilde{m}}\} = (V \cap \{1, \dots, t_{\tilde{m}}\}) \cup (U_t \cap \{1, \dots, t_{\tilde{m}}\})$. Furthermore, $U_t \cap \{1, \dots, t_{\tilde{m}}\} = U_{t_{\tilde{m}}}$.

(b) First, note that $(V \cap \{1, \dots, t_{\tilde{m}}\}) = (V \cap \{1, \dots, t_{\tilde{m}-1}\}) \cup (V \cap \{t_{\tilde{m}-1} + 1, \dots, t_{\tilde{m}}\})$. For the second part of the union, we can write $V \cap \{t_{\tilde{m}-1} + 1, \dots, t_{\tilde{m}}\} = S_t \cap \{t_{\tilde{m}-1} + 1, \dots, t_{\tilde{m}}\}$. To see this, note that $V \subseteq S_t$ and, due to the definition of $\tilde{m}$, $|V \cap \{t_{\tilde{m}-1} + 1, \dots, t_{\tilde{m}}\}| = |V \cap \{1, \dots, t_{\tilde{m}}\}| - |V \cap \{1, \dots, t_{\tilde{m}-1}\}| \geq |S_{t_{\tilde{m}}}| - |S_{t_{\tilde{m}-1}}| = |S_t \cap \{t_{\tilde{m}-1} + 1, \dots, t_{\tilde{m}}\}|$. Now consider $J = V \cap \{1, \dots, t_{\tilde{m}-1}\}$. By the definition of $\tilde{m}$, we have that $|J \cap \{1, \dots, t_i\}| \leq |S_{t_i}| - i$ for all $i \in \{1, \dots, \tilde{m} - 1\}$, with an equality for $i = \tilde{m} - 1$. Among all sets $J \subseteq S_t$ that satisfy the aforementioned constraints, the set A at step $t_{\tilde{m}-1}$ *after* completing line 13 of the algorithm, i.e. $A_{(t_{\tilde{m}-1})+}$, minimizes the product of the e-values. This is because $|(S_t \setminus J) \cap \{1, \dots, t_i\}| \geq i$, $i \in \{1, \dots, \tilde{m} - 1\}$, and the `SeqE-Guard` algorithm exactly removes those indices from S_t that correspond to the largest possible e-values. Since no indices were removed between the times $t_{\tilde{m}-1}$ and $t_{\tilde{m}}$, it follows that $A_{(t_{\tilde{m}}-1)+} \cap \{1, \dots, t_{\tilde{m}-1}\} = A_{t_{\tilde{m}}} \cap \{1, \dots, t_{\tilde{m}-1}\}$, showing the claim.

(c) Again, since no indices were removed between the times $t_{\tilde{m}-1}$ and $t_{\tilde{m}}$, we have $A_{t_{\tilde{m}}} \cap \{t_{\tilde{m}-1} + 1, \dots, t_{\tilde{m}}\} = S_t \cap \{t_{\tilde{m}-1} + 1, \dots, t_{\tilde{m}}\}$.

(d) This follows immediately from the fact that $\Pi_{A_{t_i} \cup U_{t_i}} \geq 1/\alpha$ for all $i \in \{1, \dots, m\}$.

Hence, it remains to show that $\boldsymbol{d}^{\boldsymbol{\phi}}(S_t) < d_{t-1} + 1$, if $\Pi_{A_t \cup U_t} < 1/\alpha$. Since $\Pi_{(A_t \cap \{1,\dots,i\}) \cup U_t} < 1/\alpha$ for all $i \in \{1, \dots, t-1\}$, $\Pi_{A_t \cup U_t} < 1/\alpha$ implies that $\phi_{A_t \cup U_t} = 0$. Furthermore, since $|S_t \setminus A_t| = d_{t-1}$, the claim follows. □

References

[1] 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*, 76(4):771–794, 2014.

[2] 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: Statistical Methodology*, 57(1):289–300, 1995.

[3] Donald A Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967.

[4] Boyan Duan, Aaditya Ramdas, Sivaraman Balakrishnan, and Larry Wasserman. Interactive martingale tests for the global null. *Electronic Journal of Statistics*, 14(2):4489–4551, 2020.

[5] Lasse Fischer, Marta Bofill Roig, and Werner Brannath. The online closure principle. *The Annals of Statistics*, 52(2):817–841, 2024.

[6] Dean P Foster and Robert A Stine. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(2): 429–444, 2008.

[7] David A Freedman. On tail probabilities for martingales. *The Annals of Probability*, pages 100–118, 1975.

[8] Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035–1061, 2004.

[9] Christopher R Genovese and Larry Wasserman. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476):1408–1417, 2006.

[10] Jelle J Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

[11] Jelle J Goeman, Rosa J Meijer, Thijmen JP Krebs, and Aldo Solari. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106(4): 841–856, 2019.

[12] Jelle J Goeman, Jesse Hemerik, and Aldo Solari. Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics*, 49(2):1218–1238, 2021.

[13] Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology (with discussion)*, 2024.

[14] Jesse Hemerik and Jelle J Goeman. False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):137–155, 2018.

[15] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[16] 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.

[17] Natasha A Karp, Jeremy Mason, Arthur L Beaudet, Yoav Benjamini, Lynette Bower, Robert E Braun, Steve DM Brown, Elissa J Chesler, Mary E Dickinson, Ann M Flenniken, et al. Prevalence of sexual dimorphism in mammalian phenotypic traits. *Nature communications*, 8(1):1–12, 2017.

[18] Eugene Katsevich and Aaditya Ramdas. Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *The Annals of Statistics*, 48 (6):3465–3487, 2020.

[19] Jinzhou Li, Marloes H Maathuis, and Jelle J Goeman. Simultaneous false discovery proportion bounds via knockoffs and closed testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae012, 2024.

[20] Ruth Marcus, Peritz Eric, and K Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[21] Iqraa Meah, Gilles Blanchard, and Etienne Roquain. False discovery proportion envelopes with m-consistency. *Journal of Machine Learning Research*, 25(270):1–52, 2024.

[22] Violeta Muñoz-Fuentes, Pilar Cacheiro, Terrence F Meehan, Juan Antonio Aguilar-Pimentel, Steve DM Brown, Ann M Flenniken, Paul Flicek, Antonella Galli, Hamed Haseli Mashhadi, Martin Hrabě de Angelis, et al. The international mouse phenotyping consortium (IMPC): a functional catalogue of the mammalian genome that informs conservation. *Conservation genetics*, 19(4):995–1005, 2018.

[23] Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *arXiv preprint arXiv:2410.23614*, 2024.

[24] Aaditya Ramdas, Fanny Yang, Martin J Wainwright, and Michael I Jordan. Online control of the false discovery rate with decaying memory. *Advances in Neural Information Processing Systems*, 30, 2017.

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

[26] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

[27] David S Robertson, Jan Wildenhain, Adel Javanmard, and Natasha A Karp. onlinefdr: an r package to control the false discovery rate for growing data repositories. *Bioinformatics*, 35(20): 4196–4199, 2019.

[28] David S Robertson, James MS Wason, and Aaditya Ramdas. Online multiple hypothesis testing. *Statistical Science*, 38(4):557, 2023.

[29] Joseph P Romano, Azeem Shaikh, and Michael Wolf. Consonance and the closure method in multiple testing. *The International Journal of Biostatistics*, 7(1):Art. 12, 27, 2011.

[30] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society (with discussion)*, 184(2):407–431, 2021.

[31] Eckart Sonnemann. *Allgemeine Lösungen multipler Testprobleme*. Universität Bern. Institut für Mathematische Statistik und Versicherungslehre, 1982.

[32] Samuel A Stouffer, Edward A Suchman, Leland C DeVinney, Shirley A Star, and Robin M Williams Jr. *The American soldier: Adjustment during army life (studies in social psychology in World War II)*, volume 1. Princeton Univ. Press, 1949.

[33] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

[34] Anna Vesely, Livio Finos, and Jelle J Goeman. Permutation-based true discovery guarantee by sum tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3): 664–683, 2023.

[35] Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars Paris, 1939.

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

[37] Vladimir Vovk and Ruodu Wang. Confidence and discoveries with e-values. *Statistical Science*, 38(2):329–354, 2023.

[38] Vladimir Vovk and Ruodu Wang. True and false discoveries with independent and sequential e-values. *Canadian Journal of Statistics*, 52(4):e11833, 2024.

[39] Vladimir G Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55(2):317–341, 1993.

[40] A Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117–186, 1945.

[41] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022.

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

[43] Ziyu Xu and Aaditya Ramdas. Online multiple testing with e-values. In *International Conference on Artificial Intelligence and Statistics*, pages 3997–4005. PMLR, 2024.