THEORY PROBAB. APPL.
Vol. 68, No. 1, pp. 77–87

OPTIMAL INFORMATION USAGE IN BINARY SEQUENTIAL HYPOTHESIS TESTING*

M. DÖRPINGHAUS†, I. NERI‡, É. ROLDÁN§, AND F. JÜLICHER¶

Abstract. An interesting question is whether an information theoretic interpretation can be given of optimal algorithms in sequential hypothesis testing. We prove that for the binary sequential probability ratio test of a continuous observation process, the mutual information between the observation process up to the decision time and the actual hypothesis conditioned on the decision variable is equal to zero. This result can be interpreted as an optimal usage of the information on the hypothesis available in the observations by the sequential probability ratio test. As a consequence, the mutual information between the random decision time of the sequential probability ratio test and the actual hypothesis conditioned on the decision variable is also equal to zero.

Key words. sequential hypothesis testing, sequential probability ratio test, mutual information

DOI. 10.1137/S0040585X97T991295

1. Introduction.

Sequential hypothesis tests are used to make fast and reliable decisions. These tests should use the available measurements in an optimal way such that the average time to make a decision is minimized. Binary sequential hypothesis testing was first mathematically formulated in the seminal work by Wald, who introduced the sequential probability ratio test — a particular realization of a binary sequential hypothesis test [18]. The sequential probability ratio test takes binary decisions on two hypotheses based on sequential observations of a stochastic process. The sequential probability ratio test accumulates the likelihood ratio given by the sequence of observations and makes a decision as soon as this cumulative likelihood ratio exceeds or falls below two given thresholds which depend on the required reliability of the decision. A key characteristic of such a sequential probability ratio test is that its termination time is a random quantity depending on the actual realization of the observation sequence.

For independent and identically distributed (i.i.d.) observations the sequential probability ratio test yields minimum mean decision times for decisions with a given probability of error and a given hypothesis [19]. Moreover, for continuous observation processes it was proved that the sequential probability ratio test is optimal in the sense of minimizing the Kullback–Leibler divergences between the two measures describing the statistics of the observation process up to the decision time under the two hypotheses [16, section 3.3]. The sequential probability ratio test was applied to non-i.i.d.

*Received by the editors December 7, 2020; revised July 18, 2022; published electronically May 4, 2023. This work was partly supported by the German Research Foundation (DFG) within the Cluster of Excellence EXC 1056 "Center for Advancing Electronics Dresden (cfaed)" and within the CRC 912 "Highly Adaptive Energy-Efficient Computing (HAEC)." Originally published in the Russian journal *Teoriya Veroyatnostei i ee Primeneniya*, 68 (2023), pp. 93–105.
https://doi.org/10.1137/S0040585X97T991295

†Vodafone Chair Mobile Communications Systems and Center for Advancing Electronics Dresden (cfaed), Technische Universität Dresden, Dresden, Germany (meik.doerpinghaus@tu-dresden.de).
‡Department of Mathematics, King's College London, London, UK (izaak.neri@kcl.ac.uk).
§ICTP — Abdus Salam International Centre for Theoretical Physics, Trieste, Italy (edgar@ictp.it).
¶Max-Planck-Institute for the Physics of Complex Systems, Dresden, Germany, and Center for Advancing Electronics Dresden (cfaed), Technische Universität Dresden, Dresden, Germany (julicher@pks.mpg.de).

observation processes and to nonhomogeneous and correlated continuous-time processes and was generalized for multiple hypotheses [16]. However, optimality has not been proved for most of these cases. Nevertheless, a weaker statement of asymptotic optimality (in the sense of minimum mean decision times) for the case when probabilities of errors tend to zero was proved for a broad class of stochastic processes; see, e.g., [9], [17], [15], [5], [16].

Here, we ask whether we can develop an understanding of optimality in terms of information usage of sequential hypothesis testing. In this regard, intuition lets us conjecture that an optimal decision contains all information on the actual hypothesis given by the observation process. However, this would imply that the observation process up to the decision time does not contain any additional information on the hypothesis beyond the information given by the decision itself. As a consequence, this would also mean that the random time at which a decision is made does not contain any additional information on the hypothesis beyond the information given by the decision itself. In the following we formalize these statements.

We consider a sequential probability ratio test which takes as input the realization of a *continuous* stochastic process corresponding to either hypothesis H_1 or hypothesis H_2 and gives as output a binary decision variable $\mathsf{D}_{\mathrm{w}} \in \{1,2\}$ (corresponding to hypotheses H_1 and H_2, respectively) at the random decision time T_{w} elapsed since the beginning of the observations. We show that for the sequential probability ratio test, the mutual information

$$I(\mathsf{H}; \mathsf{X}_0^{\mathsf{T}_{\mathrm{w}}} \,|\, \mathsf{D}_{\mathrm{w}}) = 0, \tag{1.1}$$

where $\mathsf{H} \in \{1,2\}$ (corresponding to hypotheses H_1 and H_2) denotes the random binary hypothesis and $\mathsf{X}_0^{\mathsf{T}_{\mathrm{w}}}$ is the observation process from time $t = 0$ until the decision time T_{w}. Equation (1.1) implies that the observation process $\mathsf{X}_0^{\mathsf{T}_{\mathrm{w}}}$ up to the decision time T_{w} does not contain any information on which hypothesis is true beyond the decision outcome D_{w}.

Condition (1.1) readily implies that

$$I(\mathsf{H}; \mathsf{T}_{\mathrm{w}} \,|\, \mathsf{D}_{\mathrm{w}}) = 0, \tag{1.2}$$

which states that the distribution of the decision time T_{w} given a certain decision outcome is independent of the actual hypothesis. In other words, equation (1.2) states that the decision time T_{w} of the sequential probability ratio test does not contain any information on which hypothesis is true beyond the decision outcome D_{w}. As a consequence, the sequential probability ratio test for continuous observation processes, which is optimal in the sense of minimizing Kullback–Leibler divergences (as stated above), minimizes the mutual information $I(\mathsf{H}; \mathsf{T}_{\mathrm{w}} \,|\, \mathsf{D}_{\mathrm{w}})$. Note that the mutual information criterion in (1.2) is not a sufficient condition for minimizing the mean decision time. This can be easily verified by adding a constant time delay t_{delay} to the actual decision time for which we still have $I(\mathsf{H}; \mathsf{T}_{\mathrm{w}} + t_{\mathrm{delay}} \,|\, \mathsf{D}_{\mathrm{w}}) = 0$.

Relation to other work. For the case of i.i.d. observations, low error probabilities α_1 and α_2 of the first and second kinds, $\alpha_1 = \alpha_2$, and equally likely hypotheses H_1 and H_2, conditions for optimal usage of information were derived in [4]. Namely, it was shown that a condition similar to $I(\mathsf{H}; \mathsf{T}_{\mathrm{w}} \,|\, \mathsf{D}_{\mathrm{w}}) = 0$ holds for a certain model for the observation process; see [4, Theorem 1]. Moreover, the relations given in Corollary 3.1 in the present paper share similarities with equalities on the first-passage-time distributions of the stochastic entropy production derived in [13], [12] and equalities for first-passage-time distributions in random walks [8]. In addition, in communication theory, relations reminiscent of Corollary 3.1 were found to show that the probability of cycle slips to the positive/negative boundary in phase-locked loops used for synchronization is independent of time [10, eq. (74)].

Notation. We denote random variables (r.v.'s) by uppercase sans serif letters, e.g., X. All random quantities are defined on the measurable space $(\Omega, \mathcal{F})$ and are governed by the probability measure $\mathbf{P}$. Mathematical expectation with respect to $\mathbf{P}$ is denoted by $\mathbf{E}(\,\cdot\,)$. For discrete r.v.'s, $\mathbf{P}_{\mathsf{Y}=y}(\mathsf{X}=x)$ denotes the probability of $\mathsf{X}=x$ given $\mathsf{Y}=y$, and $\mathbf{E}_{\mathsf{Y}=y}(\,\cdot\,)$ is the expectation conditioned on $\mathsf{Y}=y$; analogously, we use $\mathbf{P}_{\mathsf{Y}}(\mathsf{X}=x)$ for the probability of $\mathsf{X}=x$ given Y, and $\mathbf{E}_{\mathsf{Y}}(\,\cdot\,)$ for the expectation conditioned on Y. The restriction $\mathbf{P}|_{\mathcal{G}}$ of the measure $\mathbf{P}$ to a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ is defined by

$$\mathbf{P}|_{\mathcal{G}}(\Phi) = \begin{cases} \mathbf{P}(\Phi) & \text{if } \Phi \in \mathcal{G}, \\ 0 & \text{if } \Phi \in \mathcal{F} \setminus \mathcal{G}. \end{cases} \tag{1.3}$$

We write $\int_{\Phi} \mathsf{X}\, \mathrm{d}\mathbf{P}|_{\mathcal{G}}$ for an integral on the set Φ of the r.v. X over the probability measure $\mathbf{P}|_{\mathcal{G}}$. We denote the Radon–Nikodým derivative of the measure $\mathbf{P}$ with respect to the measure $\mathbf{Q}$ by $\mathrm{d}\mathbf{P}/\mathrm{d}\mathbf{Q}$. In addition, ln denotes the natural logarithm. The mutual information and the conditional mutual information are defined by $I(\mathsf{X};\mathsf{Y}) = \mathbf{E}\big(\ln(\mathrm{d}\mathbf{P}_{\mathsf{Y}}|_{\mathcal{F}(\mathsf{X})}/\mathrm{d}\mathbf{P}|_{\mathcal{F}(\mathsf{X})})\big)$ and $I(\mathsf{X};\mathsf{Y}\,|\,\mathsf{Z}) = \mathbf{E}\big(\ln(\mathrm{d}\mathbf{P}_{\mathsf{Y},\mathsf{Z}}|_{\mathcal{F}(\mathsf{X})}/\mathrm{d}\mathbf{P}_{\mathsf{Z}}|_{\mathcal{F}(\mathsf{X})})\big)$, respectively, where Y and Z are discrete r.v.'s, and $\mathcal{F}(\mathsf{X})$ is the sub-σ-algebra of $\mathcal{F}$ generated by the r.v. X.

Organization of the paper. In section 2 we describe the system setup in detail. Subsequently, in section 3 we state the main theorems and corollaries regarding the optimal information usage of the sequential probability ratio test. In section 4, we discuss the given results. Finally, the proofs are presented in the appendix.

2. System setup.

We consider a sequential binary decision problem based on an observation process X_t with continuous time index $t \in \mathbf{R}_+$. The stochastic process X_t is generated by one of two possible models corresponding to two hypotheses H_1 and H_2. To describe the statistics of the process X_t, we consider the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbf{P})$ with $\{\mathcal{F}_t\}_{t\geqslant 0}$, the natural filtration generated by the observation process X_t and the hypothesis H. We consider H to be a time independent r.v. The statistics of the observation process under the two hypotheses are described by the conditional probability measures given the hypothesis $\mathbf{P}_{\mathsf{H}=l}(\Phi) = \mathbf{E}_{\mathsf{H}=l}(1_{\Phi})$ with $l \in \{1,2\}$ corresponding to the hypotheses H_1 and H_2, respectively; here $1_{\Phi}(\omega)$ is the indicator function on the set $\Phi \in \mathcal{F}$. We assume that $\mathbf{P}(\mathsf{H}=1) > 0$ and $\mathbf{P}(\mathsf{H}=2) > 0$ and that the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbf{P})$ is complete, which means that $\mathcal{F}$ contains all sets $\Phi \subset \Omega$ for which there exist sets $\Phi_1 \in \mathcal{F}$ and $\Phi_2 \in \mathcal{F}$ such that $\Phi_1 \subset \Phi \subset \Phi_2$ and $\mathbf{P}(\Phi_2) = \mathbf{P}(\Phi_1)$, and $\mathcal{F}_0$ contains all $\Phi \in \mathcal{F}$ with $\mathbf{P}(\Phi) = 0$ [11, Chap. 1]. Here and in what follows, we use the shorthand notation $\mathbf{P}(\mathsf{H}=1) = \mathbf{P}(\{\omega \in \Omega \colon \mathsf{H}(\omega) = 1\})$ for probabilities of sets. We consider the filtration $\{\mathcal{F}_t\}_{t\geqslant 0}$ to be right-continuous [11, Chap. 1], i.e., $\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s$ for all times $t \in \mathbf{R}_+$. The two measures $\mathbf{P}_{\mathsf{H}=1}$ and $\mathbf{P}_{\mathsf{H}=2}$ are assumed to be locally mutually absolutely continuous with respect to the filtration $\{\mathcal{F}\}_{t\geqslant 0}$ [11].

A sequential hypothesis test makes binary decisions based on sequential observations of the process X_t and tries to infer which of the hypotheses H_1 and H_2 is true. A sequential hypothesis test $\delta = (\mathsf{D}, \mathsf{T})$ returns a binary output D at a random time T. The decision time $\mathsf{T} \in [0, \infty]$ is a stopping time, i.e., a random time for which $\{\omega \in \Omega \colon \mathsf{T}(\omega) \leqslant t\} \in \mathcal{F}_t$ holds for all $t \in \mathbf{R}_+$. The decision function $\mathsf{D} \in \{1, 2\}$ is an $\mathcal{F}_{\mathsf{T}}$-measurable r.v., where $\mathcal{F}_{\mathsf{T}} = \{\Phi \in \mathcal{F} \colon \Phi \cap \{\omega \in \Omega \colon \mathsf{T}(\omega) \leqslant t\} \in \mathcal{F}_t \ \forall\, t\}$.

We now consider the following set of sequential hypothesis tests with given reliabilities:

$$\begin{aligned}\mathcal{C}(\alpha_1, \alpha_2) = \big\{\delta \colon \mathbf{P}_{\mathsf{H}=1}(\mathsf{D} = 2) \leqslant \alpha_2,\, \mathbf{P}_{\mathsf{H}=2}(\mathsf{D} = 1) \leqslant \alpha_1, \\ \mathbf{E}_{\mathsf{H}=l}(\mathsf{T}) < \infty,\, l \in \{1, 2\}\big\};\end{aligned} \tag{2.1}$$

here $\mathbf{E}_{\mathsf{H}=l}(\mathsf{T})$ denotes the expected termination time in the case when hypothesis l is true, and α_1 and α_2 are the maximum allowed error probabilities of the two error types. We assume that $\alpha_1, \alpha_2 \in (0, 0.5)$. Notice that we restrict ourselves to tests which have a finite mean decision time. This assumption is fulfilled in many cases like the case of i.i.d. or stationary observation processes [16]. Note that the set of sequential hypothesis tests given by $\mathcal{C}(\alpha_1, \alpha_2)$ does not consider prior knowledge on the statistics of H.

According to [16], an optimality criterion in terms of Kullback–Leibler divergences is given by the following definition.

Definition 2.1 (optimality in terms of Kullback–Leibler divergence). *An optimal test* $\delta^* = (\mathsf{D}^*, \mathsf{T}^*) \in \mathcal{C}(\alpha_1, \alpha_2)$ *minimizes the Kullback–Leibler divergences, viz.*,

$$\inf_{\delta \in \mathcal{C}(\alpha_1, \alpha_2)} \mathbf{E}_{\mathsf{H}=1}\left(\ln \frac{\mathrm{d}\mathbf{P}_{\mathsf{H}=1}|_{\mathcal{F}_{\mathsf{T}}}}{\mathrm{d}\mathbf{P}_{\mathsf{H}=2}|_{\mathcal{F}_{\mathsf{T}}}}\right) = \mathbf{E}_{\mathsf{H}=1}\left(\ln \frac{\mathrm{d}\mathbf{P}_{\mathsf{H}=1}|_{\mathcal{F}_{\mathsf{T}^*}}}{\mathrm{d}\mathbf{P}_{\mathsf{H}=2}|_{\mathcal{F}_{\mathsf{T}^*}}}\right), \tag{2.2}$$

$$\inf_{\delta \in \mathcal{C}(\alpha_1, \alpha_2)} \mathbf{E}_{\mathsf{H}=2}\left(\ln \frac{\mathrm{d}\mathbf{P}_{\mathsf{H}=2}|_{\mathcal{F}_{\mathsf{T}}}}{\mathrm{d}\mathbf{P}_{\mathsf{H}=1}|_{\mathcal{F}_{\mathsf{T}}}}\right) = \mathbf{E}_{\mathsf{H}=2}\left(\ln \frac{\mathrm{d}\mathbf{P}_{\mathsf{H}=2}|_{\mathcal{F}_{\mathsf{T}^*}}}{\mathrm{d}\mathbf{P}_{\mathsf{H}=1}|_{\mathcal{F}_{\mathsf{T}^*}}}\right). \tag{2.3}$$

Definition 2.1 states that a test is optimal if the statistics of the observation process X_t under the two hypotheses H_1 and H_2 up to the decision time T^* are more similar (i.e., less distinguishable) than for any other tests in $\mathcal{C}(\alpha_1, \alpha_2)$.

For continuous observation processes X_t, optimality in terms of Definition 2.1 is achieved by the sequential probability ratio test $(\mathsf{D}_{\mathrm{w}}, \mathsf{T}_{\mathrm{w}}) \in \mathcal{C}(\alpha_1, \alpha_2)$, which was introduced by Wald [18], and which is known to achieve the minimum Kullback–Leibler divergence for given reliability constraints [16]. This test observes X_t until the cumulated log-likelihood ratio

$$\mathsf{S}_t = \ln \frac{\mathrm{d}\mathbf{P}_{\mathsf{H}=1}|_{\mathcal{F}_t}}{\mathrm{d}\mathbf{P}_{\mathsf{H}=2}|_{\mathcal{F}_t}}, \qquad t \geqslant 0, \tag{2.4}$$

exceeds (falls below) a prescribed threshold L_1 (L_2) for the first time. Note that $\mathsf{S}_0 = 0$. The test decides $\mathsf{D}_{\mathrm{w}} = 1$ $(\mathsf{D}_{\mathrm{w}} = 2)$, i.e., for H_1 (H_2), when S_t first crosses L_1 (L_2). The thresholds L_1 and L_2 are given by

$$L_1 = \ln \frac{1 - \alpha_2}{\alpha_1}, \tag{2.5}$$

$$L_2 = \ln \frac{\alpha_2}{1 - \alpha_1}. \tag{2.6}$$

In summary, the sequential probability ratio test decides at the time

$$\mathsf{T}_{\mathrm{w}} = \min\{t \in \mathbf{R}_+ \colon \mathsf{S}_t \notin (L_2, L_1)\} \tag{2.7}$$

for the decision

$$\mathsf{D}_{\rm w} = \begin{cases} 1 & \text{if } \mathsf{S}_{\mathsf{T}_{\rm w}} \geqslant L_1, \\ 2 & \text{if } \mathsf{S}_{\mathsf{T}_{\rm w}} \leqslant L_2. \end{cases} \tag{2.8}$$

For the sequential probability ratio test of a continuous observation process, the error probabilities are equal to the maximum allowed error probabilities as stated by the following lemma.

LEMMA 2.1. *Let* $\delta = (\mathsf{D}_{\rm w}, \mathsf{T}_{\rm w}) \in \mathcal{C}(\alpha_1, \alpha_2)$. *If* S_t *is almost surely continuous, then*

$$\mathbf{P}_{\mathsf{H}=2}(\mathsf{D}_{\rm w} = 1) = \alpha_1, \tag{2.9}$$

$$\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_{\rm w} = 2) = \alpha_2. \tag{2.10}$$

For a proof of Lemma 2.1, see the appendix.

For observation processes X_t with i.i.d. increments, optimality in the sense of Definition 2.1 also implies that the mean decision times $\mathbf{E}_{\mathsf{H}=1}(\mathsf{T})$ and $\mathbf{E}_{\mathsf{H}=2}(\mathsf{T})$ are minimized [16]. Therefore, for this particular case, optimality in the sense of Definition 2.1 is equivalent to optimality in the sense of minimizing the mean decision times. To the best of our knowledge, it is not known whether there exists a test that is optimal in the sense of minimizing mean decision times for non-i.i.d. continuous observation processes.

3. Main results.

Here we state all the main results of this paper; the proofs can be found in the appendix.

The first main result of the present paper is a symmetry relation for the probability of events in the σ-algebra $\mathcal{F}_{\mathsf{T}_{\rm w}}$ for continuous observation processes.

THEOREM 3.1. *Consider the sequential probability ratio test given by* (2.7) *and* (2.8), *which is defined on the system described in section* 2, *and let us assume that* $\mathbf{E}_{\mathsf{H}}(\mathsf{T}_{\rm w}) < \infty$ *so that* $(\mathsf{D}_{\rm w}, \mathsf{T}_{\rm w}) \in \mathcal{C}(\alpha_1, \alpha_2)$. *If* S_t *is almost surely continuous, then*

$$\mathbf{P}_{\mathsf{H}=1,\,\mathsf{D}_{\rm w}=1}(\Phi) = \mathbf{P}_{\mathsf{H}=2,\,\mathsf{D}_{\rm w}=1}(\Phi), \tag{3.1}$$

$$\mathbf{P}_{\mathsf{H}=1,\,\mathsf{D}_{\rm w}=2}(\Phi) = \mathbf{P}_{\mathsf{H}=2,\,\mathsf{D}_{\rm w}=2}(\Phi) \tag{3.2}$$

for all $\Phi \in \mathcal{F}_{\mathsf{T}_{\rm w}}$, *where* $\mathcal{F}_{\mathsf{T}_{\rm w}} = \big\{A \in \mathcal{F}\colon A \cap \{\mathsf{T}_{\rm w} \leqslant t\} \in \mathcal{F}_t \ \forall\, t\big\}$.

Theorem 3.1 implies the following corollary.

COROLLARY 3.1. *Under the same conditions as in Theorem* 3.1, *for all* $t \geqslant 0$,

$$\mathbf{P}_{\mathsf{H}=1,\,\mathsf{D}_{\rm w}=1}(\mathsf{T}_{\rm w} \leqslant t) = \mathbf{P}_{\mathsf{H}=2,\,\mathsf{D}_{\rm w}=1}(\mathsf{T}_{\rm w} \leqslant t), \tag{3.3}$$

$$\mathbf{P}_{\mathsf{H}=1,\,\mathsf{D}_{\rm w}=2}(\mathsf{T}_{\rm w} \leqslant t) = \mathbf{P}_{\mathsf{H}=2,\,\mathsf{D}_{\rm w}=2}(\mathsf{T}_{\rm w} \leqslant t). \tag{3.4}$$

In order to study the optimal information usage of the sequential probability ratio test, we consider the mutual information between the trajectory of the observation process $\mathsf{X}_0^{\mathsf{T}_{\rm w}}$ up to the decision time $\mathsf{T}_{\rm w}$ and the hypothesis H conditioned on the decision variable $\mathsf{D}_{\rm w}$, which is given by [6]

$$\begin{aligned} I(\mathsf{H}; \mathsf{X}_0^{\mathsf{T}_{\rm w}} \mid \mathsf{D}_{\rm w}) = {} & \sum_{i=1}^{2}\sum_{j=1}^{2} \mathbf{P}(\mathsf{H} = i,\, \mathsf{D}_{\rm w} = j) \\ & \times \int_{\{\omega \in \Phi\colon\, \mathsf{D}_{\rm w}(\omega) = j\}} \mathrm{d}\mathbf{P}_{\mathsf{H}=i,\,\mathsf{D}_{\rm w}=j}|_{\mathcal{F}_{\mathsf{T}_{\rm w}}} \ln\!\left(\frac{\mathrm{d}\mathbf{P}_{\mathsf{H}=i}|_{\mathcal{F}_{\mathsf{T}_{\rm w}}}}{\mathrm{d}\mathbf{P}|_{\mathcal{F}_{\mathsf{T}_{\rm w}}}} \, \frac{\mathbf{P}(\mathsf{D}_{\rm w} = j)}{\mathbf{P}_{\mathsf{H}=i}(\mathsf{D}_{\rm w} = j)} \right). \end{aligned} \tag{3.5}$$

Theorem 3.1 implies the following theorem.

THEOREM 3.2. *Under the same conditions as in Theorem* 3.1,

$$I(\mathsf{H}; \mathsf{X}_0^{\mathsf{T}_\mathrm{w}} \,|\, \mathsf{D}_\mathrm{w}) = 0. \tag{3.6}$$

Theorem 3.2 states that the sequential probability ratio test for continuous observation processes X_t makes optimal use of the available information in the following sense: The decision variable D_w contains all the information about the hypothesis H in the trajectory of the observation process X_t up to the decision time T_w.

Theorem 3.2 immediately implies the following corollary.

COROLLARY 3.2. *Under the same conditions as in Theorem* 3.1, *the following equality for mutual information holds*:

$$I(\mathsf{H}; \mathsf{T}_\mathrm{w} \,|\, \mathsf{D}_\mathrm{w}) = 0, \tag{3.7}$$

i.e., $I(\mathsf{H}; \mathsf{T}_\mathrm{w}, \mathsf{D}_\mathrm{w}) = I(\mathsf{H}; \mathsf{D}_\mathrm{w})$.

Corollary 3.2 states that in the case of optimal sequential hypothesis testing, the decision time T_w does not give any additional information on the hypothesis H beyond the decision outcome D_w. Since the mutual information is always nonnegative, this implies that the sequential probability ratio test minimizes the mutual information $I(\mathsf{H}; \mathsf{T} \,|\, \mathsf{D})$. Note that in practical cases it is much harder to measure $I(\mathsf{H}; \mathsf{X}_0^{\mathsf{T}} \,|\, \mathsf{D})$ than $I(\mathsf{H}; \mathsf{T} \,|\, \mathsf{D})$.

4. Discussion. The aim of the present paper is to give an information theoretic interpretation for optimality of sequential hypothesis testing algorithms. In this regard, the main result (Theorem 3.2) has the following appealing interpretation: At the decision time the sequential probability ratio test has exploited all information on the hypothesis available in the observation process.

Theorem 3.2 holds for continuous observation processes for which it was proved that the sequential probability ratio test is also optimal in the sense of minimizing the Kullback–Leibler divergence as stated in Definition 2.1 [16, section 3.3]. If, in addition, the increments of the observation process are i.i.d. r.v.'s, the sequential probability ratio test is optimal in the sense of minimizing the mean decision time [19], [16]. This raises the questions of whether these different optimality criteria are related to one another, and whether optimal sequential hypothesis tests in discrete time or for multiple hypotheses also make optimal use of the information in the observation process. Another interesting question for future work is whether there exist sequential hypothesis tests that satisfy (3.6) and are not sequential probability ratio tests.

The main results of this paper may also be interesting for applications. For example, Theorem 3.1 and Theorem 3.2 can be used to determine how close to optimality a given black box decision system operates. In this regard, note that $I(\mathsf{H}; \mathsf{X}_0^{\mathsf{T}} \,|\, \mathsf{D})$ can be interpreted as a measure of how close the black box decision device operates to optimality, where $I(\mathsf{H}; \mathsf{X}_0^{\mathsf{T}} \,|\, \mathsf{D}) = 0$ shows that the black box decision device optimally uses all information on the hypothesis available in the observation process. Similarly, Corollary 3.1 and Corollary 3.2 can be applied to reject the hypothesis that the decision device operates optimally. While the latter test just provides a necessary condition for optimality, it has the advantage that it does not require access to the observation process $\mathsf{X}_0^{\mathsf{T}}$. Note that in both cases the properties of the decision-making device, such as the allowed error probabilities α_1 and α_2, are not required to test optimality of a system. In the arXiv preprint [3], preliminary results on testing optimality can be found. In summary, it should be feasible to use Theorem 3.1, Theorem 3.2, Corollary 3.1, and Corollary 3.2 to test optimal information usage of real-world systems that make decisions, e.g., human decision-making [1], decisions made by animals [7], and cell fate decisions [14].

5. Appendix: Proofs.

The following result is required in the proof of Lemma 2.1.

COROLLARY 5.1 (corollary to Doob's optional stopping theorem).
If $\mathbf{P}(\mathsf{T}_{\rm w} < \infty) = 1$, *then*

$$\mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_{\mathsf{T}_{\rm w}}}\big) = 1. \tag{5.1}$$

Proof of Corollary 5.1. We decompose $\mathbf{E}_{\mathsf{H}=1}(e^{-\mathsf{S}_{\mathsf{T}_{\rm w}}})$ into three terms

$$\mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_{\mathsf{T}_{\rm w}}}\big) = \mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_{\mathsf{T}_{\rm w}\wedge t}}\big) - \mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_t} 1_{\mathsf{T}_{\rm w}>t}\big) + \mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_{\mathsf{T}_{\rm w}}} 1_{\mathsf{T}_{\rm w}>t}\big), \tag{5.2}$$

where $\mathsf{T}_{\rm w} \wedge t = \min\{\mathsf{T}_{\rm w}, t\}$. The process $e^{-\mathsf{S}_{s\wedge t}}$ with $s \in [0,t]$ satisfies $e^{-\mathsf{S}_{s\wedge t}} = \mathbf{E}_{\mathsf{X}_0^s,\,\mathsf{H}=1}(e^{-\mathsf{S}_t})$, where $\mathbf{E}_{\mathsf{X}_0^s,\,\mathsf{H}=1}(\,\cdot\,)$ is the conditional expectation with respect to the σ-algebra $\mathcal{F}_s$ and $\mathsf{H}=1$. Such martingales were called regular martingales; see [11]. Hence, we can apply Theorem 3.6 in [11], which is Doob's optional stopping theorem for regular martingales, to the martingale $e^{-\mathsf{S}_{s\wedge t}}$ and the stopping time $\mathsf{T}_{\rm w}$ to obtain

$$\mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_{\mathsf{T}_{\rm w}\wedge t}}\big) = 1. \tag{5.3}$$

Making $t \to \infty$ of (5.2), we find that

$$\mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_{\mathsf{T}_{\rm w}}}\big) = 1 - \lim_{t\to\infty} \mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_t} 1_{\mathsf{T}_{\rm w}>t}\big) + \lim_{t\to\infty} \mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_{\mathsf{T}_{\rm w}}} 1_{\mathsf{T}_{\rm w}>t}\big). \tag{5.4}$$

For the second term on the right-hand side of (5.4) we obtain

$$\lim_{t\to\infty} \mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_t} 1_{\mathsf{T}_{\rm w}>t}\big) \leqslant e^{-L_2} \lim_{t\to\infty} \mathbf{E}_{\mathsf{H}=1}(1_{\mathsf{T}_{\rm w}>t}) = 0, \tag{5.5}$$

where we used (2.7) and $\mathbf{P}(\mathsf{T}_{\rm w} < \infty) = 1$. For the last term on the right-hand side of (5.4) we used the fact that $e^{-\mathsf{S}_{\mathsf{T}_{\rm w}}} 1_{\mathsf{T}_{\rm w}>t}$ is a nonnegative monotonic decreasing sequence in t. Therefore, we can apply the monotone convergence theorem to obtain

$$\lim_{t\to\infty} \mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_{\mathsf{T}_{\rm w}}} 1_{\mathsf{T}_{\rm w}>t}\big) = \mathbf{E}_{\mathsf{H}=1}\big(e^{-\mathsf{S}_{\mathsf{T}_{\rm w}}} \lim_{t\to\infty} 1_{\mathsf{T}_{\rm w}>t}\big) = 0. \tag{5.6}$$

Using (5.5) and (5.6) in (5.4), we conclude the proof.

Proof of Lemma 2.1. Using the version of Doob's optional stopping theorem given by Corollary 5.1 on the $\mathbf{P}_{\mathsf{H}=1}$-martingale $e^{-\mathsf{S}_t}$, we obtain (5.1). Since by assumption $(\mathsf{D}_{\rm w}, \mathsf{T}_{\rm w}) \in \mathcal{C}(\alpha_1, \alpha_2)$, we have $\mathbf{P}_{\mathsf{H}=1}(\mathsf{T}_{\rm w} < \infty) = 1$, and therefore (5.1) implies that

$$\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_{\rm w} = 1) e^{-L_1} + \mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_{\rm w} = 2) e^{-L_2} = 1, \tag{5.7}$$

where we used the fact that S_t is almost surely continuous. Moreover, since $\mathbf{P}_{\mathsf{H}=1}(\mathsf{T}_{\rm w} < \infty) = 1$, we have

$$\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_{\rm w} = 1) + \mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_{\rm w} = 2) = 1. \tag{5.8}$$

Equations (5.7) and (5.8) imply that

$$\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_{\rm w} = 2) = \frac{1 - e^{-L_1}}{e^{-L_2} - e^{-L_1}}. \tag{5.9}$$

Using (2.5) and (2.6) in (5.9), we obtain (2.10).

Analogously, using Doob's optional stopping theorem on the $\mathbf{P}_{\mathsf{H}=2}$-martingale e^{S_t}, we obtain (2.9). Lemma 2.1 is proved.

Proof of Theorem 3.1. Let $\Phi \in \mathcal{F}_{\mathsf{T}_\mathrm{w}}$. Then

$$\mathbf{P}_{\mathsf{H}=1,\,\mathsf{D}_\mathrm{w}=1}(\Phi) = \frac{\int_{\{\omega\in\Phi:\ \mathsf{D}_\mathrm{w}(\omega)=1\}} \mathrm{d}\mathbf{P}_{\mathsf{H}=1}}{\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_\mathrm{w}=1)} \tag{5.10}$$

$$= \frac{\int_{\{\omega\in\Phi:\ \mathsf{D}_\mathrm{w}(\omega)=1\}} \mathrm{d}\mathbf{P}_{\mathsf{H}=1}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}}{\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_\mathrm{w}=1)} \tag{5.11}$$

$$= \frac{\int_{\{\omega\in\Phi:\ \mathsf{D}_\mathrm{w}(\omega)=1\}} e^{\mathsf{S}_{\mathsf{T}_\mathrm{w}}}\,\mathrm{d}\mathbf{P}_{\mathsf{H}=2}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}}{\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_\mathrm{w}=1)} \tag{5.12}$$

$$= e^{L_1}\frac{\int_{\{\omega\in\Phi:\ \mathsf{D}_\mathrm{w}(\omega)=1\}} \mathrm{d}\mathbf{P}_{\mathsf{H}=2}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}}{\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_\mathrm{w}=1)} \tag{5.13}$$

$$= e^{L_1}\frac{\mathbf{P}_{\mathsf{H}=2}(\{\omega\in\Phi:\ \mathsf{D}_\mathrm{w}(\omega)=1\})}{\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_\mathrm{w}=1)} \tag{5.14}$$

$$= e^{L_1}\frac{\mathbf{P}_{\mathsf{H}=2}(\mathsf{D}_\mathrm{w}=1)}{\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_\mathrm{w}=1)}\mathbf{P}_{\mathsf{H}=2,\,\mathsf{D}_\mathrm{w}=1}(\Phi), \tag{5.15}$$

where for (5.10) and (5.15) we used Bayes's theorem, and (5.11) and (5.14) follow from

$$\mathbf{P}_{\mathsf{H}=i}\big(\{\omega\in\Phi:\ \mathsf{D}_\mathrm{w}(\omega)=1\}\big) = \mathbf{P}_{\mathsf{H}=i}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}\big(\{\omega\in\Phi:\ \mathsf{D}_\mathrm{w}(\omega)=1\}\big) \tag{5.16}$$

($i\in\{1,2\}$), which is true because of the definition of $\mathbf{P}_{\mathsf{H}=i}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}$. Moreover, for (5.12) we used the Radon–Nikodým theorem, the definition (2.4), the assumption that $\mathbf{P}_{\mathsf{H}=1}$ and $\mathbf{P}_{\mathsf{H}=2}$ are locally mutually absolutely continuous, and the assumption that $(\mathsf{D}_\mathrm{w},\mathsf{T}_\mathrm{w})\in\mathcal{C}(\alpha_1,\alpha_2)$ such that $\mathbf{P}_{\mathsf{H}=i}(\mathsf{T}_\mathrm{w}<\infty)=1$. For (5.13) we used the fact that e^{S_t} is with probability one a continuous process and achieves the value e^{L_1} at time T_w when $\mathsf{D}_\mathrm{w}=1$.

Using Lemma 2.1, (2.5), and (5.8), we have

$$\frac{\mathbf{P}_{\mathsf{H}=2}(\mathsf{D}_\mathrm{w}=1)}{\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_\mathrm{w}=1)} = e^{-L_1}. \tag{5.17}$$

Substituting (5.17) into (5.15) proves (3.1). A similar analysis verifies (3.2), and Theorem 3.1 is proved.

Proof of Corollary 3.1. Let

$$\Xi(t) = \{\omega\in\Omega:\ \mathsf{T}_\mathrm{w}(\omega)\leqslant t\} \tag{5.18}$$

be the set of trajectories for which the decision time does not exceed t. Since $\Xi(t)\in\mathcal{F}_{\mathsf{T}_\mathrm{w}}$, Theorem 3.1 applies to give

$$\mathbf{P}_{\mathsf{H}=1,\,\mathsf{D}_\mathrm{w}=1}(\Xi(t)) = \mathbf{P}_{\mathsf{H}=2,\,\mathsf{D}_\mathrm{w}=1}(\Xi(t)). \tag{5.19}$$

The probability of the set $\Xi(t)$ with respect to the measure $\mathbf{P}_{\mathsf{H}=1}$ or $\mathbf{P}_{\mathsf{H}=2}$ is equal to the cumulative distribution of the decision time T_w conditioned on the hypothesis $\mathsf{H}=1$ or $\mathsf{H}=2$, respectively, i.e.,

$$\mathbf{P}_{\mathsf{H}=1,\,\mathsf{D}_\mathrm{w}=1}(\Xi(t)) = \mathbf{P}_{\mathsf{H}=1,\,\mathsf{D}_\mathrm{w}=1}(\mathsf{T}_\mathrm{w}\leqslant t), \tag{5.20}$$

$$\mathbf{P}_{\mathsf{H}=2,\,\mathsf{D}_\mathrm{w}=1}(\Xi(t)) = \mathbf{P}_{\mathsf{H}=2,\,\mathsf{D}_\mathrm{w}=1}(\mathsf{T}_\mathrm{w}\leqslant t). \tag{5.21}$$

Combining (5.19), (5.20), and (5.21) proves (3.3). Equation (3.4) can be shown similarly. This concludes the proof of Corollary 3.1.

Proof of Theorem 3.2. The mutual information $I(\mathsf{H}; \mathsf{X}_0^{\mathsf{T}_\mathrm{w}} \mid \mathsf{D}_\mathrm{w})$ can be rewritten as

$$\begin{aligned} I(\mathsf{H}; \mathsf{X}_0^{\mathsf{T}_\mathrm{w}} \mid \mathsf{D}_\mathrm{w}) &= \sum_{i=1}^{2}\sum_{j=1}^{2} \mathbf{P}(\mathsf{H}=i,\, \mathsf{D}_\mathrm{w}=j) \\ &\quad \times \int_{\{\omega\in\Omega:\, \mathsf{D}_\mathrm{w}(\omega)=j\}} \mathrm{d}\mathbf{P}_{\mathsf{H}=i,\,\mathsf{D}_\mathrm{w}=j}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}} \ln\left(\frac{\mathbf{P}(\mathsf{D}_\mathrm{w}=j)}{\mathbf{P}_{\mathsf{H}=i}(\mathsf{D}_\mathrm{w}=j)} \frac{\mathrm{d}\mathbf{P}_{\mathsf{H}=i}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}}{\mathrm{d}\mathbf{P}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}}\right) \\ &= \sum_{i=1}^{2}\sum_{j=1}^{2} \mathbf{P}(\mathsf{H}=i,\, \mathsf{D}_\mathrm{w}=j) \int_{\{\omega\in\Omega:\, \mathsf{D}_\mathrm{w}(\omega)=j\}} \mathrm{d}\mathbf{P}_{\mathsf{H}=i,\,\mathsf{D}_\mathrm{w}=j}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}} \ln(N_{ij}(\omega)), \qquad (5.22) \end{aligned}$$

and so the argument of the logarithm in (5.22) can be expressed as

$$N_{ij} = \frac{\mathbf{P}(\mathsf{D}_\mathrm{w}=j)}{\mathbf{P}_{\mathsf{H}=i}(\mathsf{D}_\mathrm{w}=j)} \frac{\mathrm{d}\mathbf{P}_{\mathsf{H}=i}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}}{\mathrm{d}\mathbf{P}_{\mathsf{H}=1}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}\mathbf{P}(\mathsf{H}=1) + \mathrm{d}\mathbf{P}_{\mathsf{H}=2}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}\mathbf{P}(\mathsf{H}=2)}. \qquad (5.23)$$

Theorem 3.1 implies that, for all $\Phi \in \mathcal{F}_{\mathsf{T}_\mathrm{w}}$ for which $\Phi \subseteq \{\omega\in\Omega:\, \mathsf{D}_\mathrm{w}(\omega)=j\}$,

$$\frac{\mathbf{P}_{\mathsf{H}=2}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}(\Phi)}{\mathbf{P}_{\mathsf{H}=2}(\mathsf{D}_\mathrm{w}=j)} = \frac{\mathbf{P}_{\mathsf{H}=1}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}(\Phi)}{\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_\mathrm{w}=j)}. \qquad (5.24)$$

Thus, for $i=1$, N_{ij} can be expressed as

$$\begin{aligned} N_{1j} &= \frac{\mathbf{P}(\mathsf{D}_\mathrm{w}=j)}{\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_\mathrm{w}=j)} \\ &\quad \times \frac{\mathrm{d}\mathbf{P}_{\mathsf{H}=1}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}}{\mathrm{d}\mathbf{P}_{\mathsf{H}=1}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}\mathbf{P}(\mathsf{H}=1) + (\mathrm{d}\mathbf{P}_{\mathsf{H}=1}|_{\mathcal{F}_{\mathsf{T}_\mathrm{w}}}/\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_\mathrm{w}=j))\mathbf{P}_{\mathsf{H}=2}(\mathsf{D}_\mathrm{w}=j)\mathbf{P}(\mathsf{H}=2)} \\ &= \frac{\mathbf{P}(\mathsf{D}_\mathrm{w}=j)}{\mathbf{P}_{\mathsf{H}=1}(\mathsf{D}_\mathrm{w}=j)\mathbf{P}(\mathsf{H}=1) + \mathbf{P}_{\mathsf{H}=2}(\mathsf{D}_\mathrm{w}=j)\mathbf{P}(\mathsf{H}=2)} \\ &= 1. \qquad (5.25) \end{aligned}$$

Analogously, it can be shown that $N_{2j}=1$, and hence

$$I(\mathsf{H}; \mathsf{X}_0^{\mathsf{T}_\mathrm{w}} \mid \mathsf{D}_\mathrm{w}) = 0, \qquad (5.26)$$

which concludes the proof.

Proof of Corollary 3.2. Note that $\mathcal{F}(\mathsf{T}_\mathrm{w})$ is a sub-σ-algebra of $\mathcal{F}_{\mathsf{T}_\mathrm{w}}$. Now an appeal to Theorem 3.1 shows that $\mathbf{P}_{\mathsf{D}_\mathrm{w},\,\mathsf{H}=1}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})} = \mathbf{P}_{\mathsf{D}_\mathrm{w},\,\mathsf{H}=2}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}$. Hence $I(\mathsf{H}; \mathsf{T}_\mathrm{w}|\mathsf{D}_\mathrm{w})$ can be rewritten as

$$\begin{aligned} I(\mathsf{H}; \mathsf{T}_\mathrm{w} \mid \mathsf{D}_\mathrm{w}) &= \mathbf{E}\left(\ln\left(\frac{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\mathsf{H}}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w}}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}\right)\right) \\ &= -\mathbf{E}\left(\ln\left(\mathbf{P}_{\mathsf{D}_\mathrm{w}}(\mathsf{H}=1)\frac{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\,\mathsf{H}=1}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\mathsf{H}}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}} + \mathbf{P}_{\mathsf{D}_\mathrm{w}}(\mathsf{H}=2)\frac{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\,\mathsf{H}=2}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\mathsf{H}}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}\right)\right) \\ &= -\mathbf{E}\left(\ln\left(\mathbf{P}_{\mathsf{D}_\mathrm{w}}(\mathsf{H}=1)\frac{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\,\mathsf{H}=1}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\mathsf{H}}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}} + \mathbf{P}_{\mathsf{D}_\mathrm{w}}(\mathsf{H}=2)\frac{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\,\mathsf{H}=1}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\mathsf{H}}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}\right)\right) \qquad (5.27) \\ &= -\mathbf{E}\left(\ln\left(\frac{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\,\mathsf{H}=1}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\mathsf{H}}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}\right)\right) \end{aligned}$$

$$= \mathbf{P}(\mathsf{H}=1)\mathbf{E}\left(\ln\left(\frac{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\,\mathsf{H}=1}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\,\mathsf{H}=1}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}\right)\right)$$

$$+ \mathbf{P}(\mathsf{H}=2)\mathbf{E}\left(\ln\left(\frac{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\,\mathsf{H}=2}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}{\mathrm{d}\mathbf{P}_{\mathsf{D}_\mathrm{w},\,\mathsf{H}=1}|_{\mathcal{F}(\mathsf{T}_\mathrm{w})}}\right)\right)$$

$$= 0, \tag{5.28}$$

where we used Corollary 3.1 for (5.27) and (5.28). Recall that $\mathcal{F}(\mathsf{T}_\mathrm{w})$ is the sub-σ-algebra generated by T_w, whereas $\mathcal{F}_{\mathsf{T}_\mathrm{w}} = \{\Phi \in \mathcal{F} \colon \Phi \cap \{\omega \in \Omega \colon \mathsf{T}_\mathrm{w}(\omega) \leqslant t\} \in \mathcal{F}_t \ \forall\, t\}$. Corollary 3.2 is proved.

An alternative way to prove Corollary 3.2 is to use Theorem 3.2 and the data processing inequality [2].

Acknowledgments. We acknowledge Heinrich Meyr, Yannis Kalaidzidis, Mostafa Khalili-Marandi, and Marino Zerial for fruitful discussions. In addition, we thank Yan Fyodorov and Ivan Tyukin for proofreading the Russian version of the abstract.

REFERENCES

[1] R. Bogacz and K. Gurney, *The basal ganglia and cortex implement optimal decision making between alternative actions*, Neural Comput., 19 (2007), pp. 442–477, https://doi.org/10.1162/neco.2007.19.2.442.
[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006, https://doi.org/10.1002/047174882X.
[3] M. Dörpinghaus, I. Neri, É. Roldán, H. Meyr, and F. Jülicher, *Testing Optimality of Sequential Decision-Making*, preprint, https://arxiv.org/abs/1801.01574, 2018.
[4] M. Dörpinghaus, É. Roldán, I. Neri, H. Meyr, and F. Jülicher, *An information theoretic analysis of sequential decision-making*, in Proc. 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, IEEE, 2017, pp. 3050–3054, https://doi.org/10.1109/ISIT.2017.8007090.
[5] V. P. Draglia, A. G. Tartakovsky, and V. V. Veeravalli, *Multihypothesis sequential probability ratio tests* I: *Asymptotic optimality*, IEEE Trans. Inform. Theory, 45 (1999), pp. 2448–2461, https://doi.org/10.1109/18.796383.
[6] T. T. Kadota and A. D. Wyner, *Coding theorem for stationary, asymptotically memoryless, continuous-time channels*, Ann. Math. Statist., 43 (1972), pp. 1603–1611, https://doi.org/10.1214/aoms/1177692392.
[7] S. Kira, T. Yang, and M. N. Shadlen, *A neural implementation of Wald's sequential probability ratio test*, Neuron, 85 (2015), pp. 861–873, https://doi.org/10.1016/j.neuron.2015.01.007.
[8] P. L. Krapivsky and S. Redner, *First-passage duality*, J. Stat. Mech., 2018 (2018), 093208, https://doi.org/10.1088/1742-5468/aaddb3.
[9] T. L. Lai, *Asymptotic optimality of invariant sequential probability ratio tests*, Ann. Statist., 9 (1981), pp. 318–333, https://doi.org/10.1214/aos/1176345398.
[10] W. Lindsey and H. Meyr, *Complete statistical description of the phase-error process generated by correlative tracking systems*, IEEE Trans. Inform. Theory, 23 (1977), pp. 194–202, https://doi.org/10.1109/TIT.1977.1055692.
[11] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes*, *Vol.* I: *General Theory*, Appl. Math. (N.Y.) 5, Springer-Verlag, New York, 1977, https://doi.org/10.1007/978-1-4757-1665-8.
[12] I. Neri, É. Roldán, and F. Jülicher, *Statistics of infima and stopping times of entropy production and applications to active molecular processes*, Phys. Rev. X, 7 (2017), 011019, https://doi.org/10.1103/PhysRevX.7.011019.
[13] É. Roldán, I. Neri, M. Dörpinghaus, H. Meyr, and F. Jülicher, *Decision making in the arrow of time*, Phys. Rev. Lett., 115 (2015), 250602, https://doi.org/10.1103/PhysRevLett.115.250602.
[14] E. D. Siggia and M. Vergassola, *Decisions on the fly in cellular sensory systems*, Proc. Natl. Acad. Sci. USA, 110 (2013), pp. E3704–E3712, https://doi.org/10.1073/pnas.1314081110.

[15] A. G. Tartakovsky, *Asymptotic optimality of certain multihypothesis sequential tests: Non-i.i.d. case*, Stat. Inference Stoch. Process., 1 (1998), pp. 265–295, https://doi.org/10.1023/A:1009952514505.

[16] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential Analysis. Hypothesis Testing and Changepoint Detection*, Monogr. Statist. Appl. Probab. 136, CRC Press, Boca Raton, FL, 2015, https://doi.org/10.1201/b17279.

[17] A. Tartakovsky, *Asymptotically optimal sequential tests for nonhomogeneous processes*, Sequential Anal., 17 (1998), pp. 33–61, https://doi.org/10.1080/07474949808836397.

[18] A. Wald, *Sequential tests of statistical hypotheses*, Ann. Math. Statist., 16 (1945), pp. 117–186, https://doi.org/10.1214/aoms/1177731118.

[19] A. Wald and J. Wolfowitz, *Optimum character of the sequential probability ratio test*, Ann. Math. Statist., 19 (1948), pp. 326–339, https://doi.org/10.1214/aoms/1177730197.