Detecting Object Tracking Failure via Sequential Hypothesis Testing

— Supplementary Material —

A. Background knowledge about Hypothesis Testing and E-processes

A.1. Hypothesis Testing

Hypothesis testing is the statistical method we adopt to provide formal guarantees in object tracking. This section introduces the foundational concepts of hypothesis testing and p-values, and then presents the more recent e-value framework as a first step to develop e-processes in the next section. The definitions presented in this section are adapted from [33, 49].

A.1.1. Classical Hypothesis Testing and p-values

Hypothesis testing [14] is a fundamental statistical method used to make inferences about populational distributions based on sample data. It is a widely used technique in scientific research, quality control, and decision-making processes to draw conclusions based on sample evidence.

The process involves formulating two competing hypotheses:

- The **null hypothesis** (H_0): a statement that there is no effect or no difference, and any observed variation is due to random chance.
- The **alternative hypothesis** (H_1): a statement that contradicts the null hypothesis, suggesting that there is a true effect or difference.

The goal of hypothesis testing is to assess whether the observed data provides sufficient evidence to reject the null hypothesis in favor of the alternative. This is typically done by calculating a **test statistic** (a function of the observed data) and comparing it to a threshold determined by a pre-specified **significance level** (α), often set at 0.01, 0.05 or 0.1.

P-values are a widely used measure of statistical evidence in hypothesis testing. They are defined as the probability of observing a statistic as extreme as the one obtained, assuming the null hypothesis H_0 is true. Therefore, **a small p-value indicates strong evidence against** H_0. In practice, the decision of whether H_0 should be rejected is made by comparing the p-value to a predefined significance level (commonly $\alpha = 0.01$, 0.05, or 0.1). If the p-value is lower than α, the result is considered statistically significant. Conversely, a large p-value suggests that the observed data is consistent with H_0, and we fail to reject it (note that rejecting H_0 does not imply H_1 is true — it simply means we lack sufficient evidence against H_0).

Example A.1 outlines a basic hypothesis test on the mean of a normal distribution.

Example A.1 *We observe n i.i.d. samples $X_1, X_2, \ldots, X_n$ from a normal distribution with unknown mean μ and known variance σ^2. We want to test whether the mean is equal to a specified value μ_0:*

$$H_0 : \mu = \mu_0; \quad H_1 : \mu \neq \mu_0.$$

Under H_0, the sample mean is normally distributed:

$$\bar{X}_n \sim \mathcal{N}(\mu_0, \sigma^2/n).$$

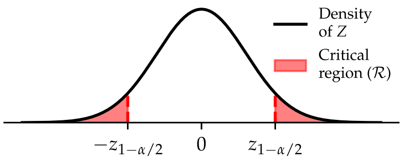
We define the standardized test statistic:

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

Given a significance level $\alpha \in (0, 1)$ (commonly $\alpha = 0.05$), we reject the null hypothesis if the test statistic falls in the critical region:

$$\mathcal{R} = \{|Z| > z_{1-\alpha/2}\},$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution.



In other words, ***we reject the null if the statistic falls in the most extreme α% fraction of its distribution under the null****.*
Alternatively, we compute the p-value:

$$p = 2 \cdot \mathbb{P}(Z > |z|),$$

where z is the observed value of the statistic, and reject H_0 if $p < \alpha$.

Let's now formalize the hypothesis testing framework. We begin by defining hypotheses as sets of probability measures:

Definition A.2 *Given a measurable space $(\Omega, \mathcal{F})$, a* ***hypothesis*** *is a set of probability measures on $(\Omega, \mathcal{F})$.*
Usually, we use the following notation for a hypothesis test:

$$H_0 : \mathbb{P} \in \mathcal{P}; \quad H_1 : \mathbb{Q} \in \mathcal{Q}, \tag{16}$$

with $\mathcal{P} \cap \mathcal{Q} = \varnothing$.
We say that a hypothesis is ***simple*** *if $|\mathcal{P}| = 1$ and* ***composite*** *otherwise.*

To evaluate evidence in data, we define tests:

Definition A.3 *A binary test ϕ is a $\{0,1\}-$valued random variable.*

If the test takes the value $\phi = 1$, we reject the null hypothesis, while $\phi = 0$ means not rejecting it. We can compute the **Type-I error** or **false positive rate** of a test ϕ for a distribution $\mathbb{P} \in \mathcal{P}$ as $\mathbb{E}_{\mathbb{P}}[\phi]$. That is, the probability of rejecting the null hypothesis when it is actually true. For a composite hypothesis $\mathcal{P}$, we say that the test has level α if its Type-I error is at most α for every $\mathbb{P} \in \mathcal{P}$.

Let's now formalize the concept of p-values.

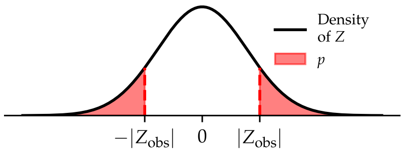
Definition A.4 *A* ***p-variable*** *P for $\mathcal{P}$ is a $[0, \infty)-$ valued random variable such that $\mathbb{P}(P \leqslant \alpha) \leqslant \alpha$ for all $\alpha \in (0,1)$ and all $\mathbb{P} \in \mathcal{P}$.*
- *If the equality holds in the condition above ($\mathbb{P}(P \leqslant \alpha) = \alpha$), we say that P is an* ***exact*** *p-variable.*

We use the term **p-value** to refer to the realized value of a p-variable computed from observed data[6]. Intuitively, for usual significance levels α, small p-values are unlikely under the null hypothesis. Therefore, **a low p-value gives statistical evidence against** H_0, while a high p-value indicates that the data is consistent with H_0, and there is insufficient evidence to reject it.

In practice, p-variables are usually computed as the cumulative distribution function of a test statistic, as shown in.

Example A.5 *Let's go back to Example A.1 to better understand how the p-variable is constructed. We defined the p-variable P as the probability that a standard normal variable exceeds the observed test statistic in magnitude:*

$$P = 2 \cdot \mathbb{P}(Z > |Z_{obs}|), \quad Z \sim \mathcal{N}(0,1).$$



Note that P is a function of the test statistic Z. Since Z_{obs} under H_0 also follows $\mathcal{N}(0,1)$, we have:

$$\begin{aligned} &\mathbb{P}_{H_0}(P \leqslant \alpha) \\ &\quad = \mathbb{P}_{H_0}(2 \cdot \mathbb{P}(Z > |Z_{obs}|) \leqslant \alpha) \\ &\quad = \mathbb{P}_{H_0}\left(|Z_{obs}| \geqslant z_{1-\alpha/2}\right) \\ &\quad = \alpha. \end{aligned}$$

Hence, P is an exact *p-variable.*

[6] In practice, "p-value" is used for both concepts, even though this is an abuse of notation.

A.1.2. Hypothesis Testing with e-values

Opposed to the classical p-value framework, hypothesis testing can also be performed using **e-values** [33], which offer a different and often more flexible approach to measuring evidence against a null hypothesis. E-values are built on expected values rather than tail probabilities. While p-values quantify how extreme the data are under the null, e-values quantify how much evidence the data provide for the alternative hypothesis, relative to the null.

We can formally define an e-variable as follows:

Definition A.6 *An* ***e-variable*** E *for* $\mathcal{P}$ *is a* $[0, \infty]-$ *valued random variable such that* $\mathbb{E}_{\mathbb{P}}\left[E\right] \leqslant 1$ *for all* $\mathbb{P} \in \mathcal{P}$.

Contrary to p-variables, e-variables are expected to be small under the null. Therefore, high e-values indicate statistical evidence against H_0.

Example A.7 shows a concrete construction of an e-variable in the Gaussian setting, using a likelihood ratio between simple hypotheses.

Example A.7 *We consider the same data as in Example A.1, where we observe* n *i.i.d. samples from a Gaussian distribution with known variance* σ^2*, but we now aim to test:*

$$H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1,$$

for some fixed alternative $\mu_1 \neq \mu_0$*. Instead of a p-variable, we construct an* ***e-variable*** *using a likelihood ratio between the alternative* μ_1 *and the null* μ_0*:*

$$\begin{aligned} E &= \frac{f(X_1, \ldots, X_n; \mu_1)}{f(X_1, \ldots, X_n; \mu_0)} \\ &= \prod_{i=1}^{n} \frac{f(X_i; \mu_1)}{f(X_i; \mu_0)} \\ &= \exp\left(\frac{1}{2\sigma^2}{}_{i=1}^{n} \left[(X_i - \mu_0)^2 - (X_i - \mu_1)^2 \right] \right) \\ &= \exp\left(\frac{n}{2\sigma^2} \left[2(\mu_1 - \mu_0)\bar{X}_n + (\mu_0^2 - \mu_1^2) \right] \right) \end{aligned}$$

This ratio compares the likelihood under a fixed alternative μ_1 *to that under the null* μ_0*. It satisfies the condition to be an (exact) e-variable:*

$$\begin{aligned} \mathbb{E}_{H_0}[E] &= \mathbb{E}_{\mu_0}\left[\frac{f(\mathbf{x}; \mu_1)}{f(\mathbf{x}; \mu_0)} \right] \\ &= \int \frac{f(\mathbf{x}; \mu_1)}{f(\mathbf{x}; \mu_0)} f(\mathbf{x}; \mu_0)\, d\mathbf{x} \\ &= \int f(\mathbf{x}; \mu_1)\, d\mathbf{x} = 1. \end{aligned}$$

We reject H_0 *if the e-variable exceeds a threshold* $1/\alpha$*:*

$$E > \frac{1}{\alpha},$$

which yields a test that controls the Type-I error at level α*.*

While classical hypothesis testing provides a rigorous framework for decision-making under uncertainty, it typically assumes a fixed dataset and does not account for scenarios where data arrives over time. However, many real-world applications such as object tracking require decisions to be made sequentially as new observations are collected. To address this, we now turn to sequential hypothesis testing, which extends classical methods to dynamic, time-evolving settings while maintaining statistical validity at each step.

A.2. Sequential Hypothesis Testing

In classical hypothesis testing, the number of samples is fixed in advance: we collect a dataset of size n, compute a test statistic, and make a binary decision to reject or retain the null hypothesis. However, in many real-world settings such as streaming data, online learning, or adaptive systems like object tracking this assumption does not hold. In sequential hypothesis testing, we receive data one point at a time and want to decide *on the fly* whether we have enough evidence to reject the null or whether we should keep receiving data.

Sequential testing introduces a fundamental challenge: if we monitor a test statistic such as a p-value or e-value repeatedly over time and stop the test when it appears significant, we risk inflating the probability of false discoveries. This issue arises because classical p-values and fixed-time e-values are not generally valid when the stopping time is data-dependent.

To address this, we must use tools designed for sequential settings. In particular, e-processes (sequences of e-values structured as nonnegative supermartingales) offer a robust solution. They remain valid under arbitrary stopping rules and allow continuous monitoring while preserving rigorous Type-I error guarantees. In this section, we introduce the basic mathematical machinery behind these tools. The definitions presented in this section are adapted from [33].

A.2.1. Random Processes and Martingales

To define sequential tests, we work with sequences of random variables that evolve over time. These are called **random processes** and are defined as follows:

Definition A.8 *We consider a sample space Ω and a filtration $\mathbf{F} = (\mathcal{F}_t)_{t \geqslant 0}$, representing the information available up to each time t. A sequence of random variables $(W_t)_{t \geqslant 0}$ is a **process** adapted to $\mathbf{F}$ if W_t is measurable with respect to $\mathcal{F}_t$ for all t. Usually, we take $\mathcal{F}_t$ to be the natural filtration of the observed data: $\mathcal{F}_t = \sigma(W_1, ..., W_t)$.*

There is a special type of random process that is relevant for our purpose: **martingales**. These can be thought of as fair game processes, where the expected future value is equal to the current one, given the past.

Definition A.9 *A process $\{W_t\}_{t \geqslant 0}$ is a **martingale** with respect to a filtration $(\mathcal{F}_t)$ if:*
(i) $\mathbb{E}[W_t] < \infty$ for all t;
(ii) $\mathbb{E}[W_t \mid \mathcal{F}_{t-1}] = X_{t-1}$ for all $t \geqslant 1$.
*If condition (2.) holds with "$\leqslant$" instead of "$=$", we say that W_t is a **supermartingale**.*

Supermartingales play a central role in safe anytime-valid inference (SAVI) because they enable the construction of **e-processes**: sequences of e-values that remain valid under arbitrary stopping.

In sequential analysis, we may want to stop the test early if we accumulate strong evidence, or continue testing if evidence is weak. To formalize this, we introduce the notion of a **stopping time**:

Definition A.10 *A **stopping time** (or **stopping rule**) τ is a nonnegative, integer-valued random variable such that the event $\{\tau \leqslant t\}$ is measurable with respect to $\mathcal{F}_t$ for all $t \geqslant 0$. That is, at time t, we can make a decision whether we should stop or continue.*

From now on, we denote the set of all stopping times as $\mathcal{T}$.

In traditional sequential tests (like Wald's Sequential Probability Ratio Test [47]), stopping rules must be predefined. However, e-processes remain valid even if we pick τ adaptively based on the data.

A.2.2. Sequential Tests

Let's now formalize the notion of sequential test

Definition A.11 *A level-α **sequential test** for $\mathcal{P}$ is a binary process ϕ such that*

$$\mathbb{P}\left(\exists t \geqslant 1 : \phi_t = 1\right) \leqslant \alpha \text{ for all } \mathbb{P} \in \mathcal{P}. \tag{17}$$

We can also identify a test by the stopping time $\tau := \inf\{t \geqslant 1 : \phi_t = 1\}$ (with the convention $\inf \varnothing = \infty$). Then, the condition in Equation Eq. 17 turns into

$$\mathbb{E}_{\mathbb{P}}\left[\phi_\tau\right] \leqslant 1 \text{ for all } \forall \mathbb{P} \in \mathcal{P}, \tau \in \mathcal{T} \tag{18}$$

In other words, a sequential test is a decision rule that can monitor data sequentially and raise an alarm (i.e., reject the null) at any point in time, but still guarantees that the probability of making a false discovery remains below α, no matter when we choose to stop. This is a much stronger requirement than in classical hypothesis testing, because it ensures validity even if the stopping time is random, data-dependent, or not specified in advance.

A.2.3. E-processes

E-processes provide such guarantees by generalizing the concept of e-values to entire processes:

Definition A.12 *A non-negative process $\{W_t\}_{t \in \mathcal{T}}$ that is adapted to $\mathcal{F}$ is called an* ***e-process*** *if*

$$\mathbb{E}_{\mathbb{P}}\left[W_\tau\right] \leqslant 1 \textit{ for any stopping time } \tau \in \mathcal{T} \textit{ and any } \mathbb{P} \in \mathcal{P}. \tag{19}$$

This definition ensures that the process controls Type-I error uniformly over time. In particular, if the null hypothesis holds, then the e-process cannot grow too large on average, regardless of when we choose to stop and make a decision.

Now, we will introduce a relevant result about e-processes that allows to build sequential tests from them with statistical guarantees on the Type-I error [44]:

Proposition A.13 ***(Ville's inequality)*** *If $\{W_t\}_{t \in}$ is an E-process for a null $\mathcal{P}$, then for any $\alpha \geqslant 1$:*

$$\sup_{P \in \mathcal{P}} P\left(\sup_t W_t \geqslant \frac{1}{\alpha}\right) \leqslant \alpha \quad \forall \alpha \in (0,1). \tag{20}$$

Although we are not including the proof of this result, note that it is just a generalization to e-processes of Markov's inequality applied to e-variables Given an e-variable E, applying Markov's inequality and the definition of e-value gives:

$$\mathbb{P}\left(E \geqslant \alpha\right) \leqslant \frac{\mathbb{E}\left[E\right]}{\alpha} \leqslant \frac{1}{\alpha}.$$

This result allows us to obtain a sequential test from an e-process by rejecting the null as soon as the process exceeds $\frac{1}{\alpha}$ *at any time*, with the guarantee that Type-I error is upper bounded by α:

$$\tau = \inf\left\{t : W_t \geqslant \frac{1}{\alpha}\right\} \tag{21}$$

To apply sequential hypothesis testing in practice, we require a principled method for constructing e-processes—test statistics that grow when the data is incompatible with the null and remain controlled otherwise. While the definition of an e-process ensures validity under arbitrary stopping, it does not prescribe how to design such a process. Indeed, this is the key challenge driving much of the recent work: how to construct suitable e-processes for increasingly complex composite hypotheses, particularly in the context of deep learning and adaptive models. In the following section, we present an approach for building e-processes by modeling statistical inference as a dynamic betting game.

A.2.4. Constructing e-processes through Testing by Betting

E-processes can be interpreted as a betting game between a statistician and nature [33, 34, 50]. The statistician starts with an initial wealth and sequentially places bets $B_t : \mathcal{W} \to [0, \infty]$ based on past observations, such that the expected value under the null is at most one:

$$\mathbb{E}_{\mathbb{P}}\left[B_t | W_1, ..., W_{t-1}\right] \leqslant 1 \quad \forall \mathbb{P} \in \mathcal{P}. \tag{22}$$

The wealth process is initialized with $W_0 = 1$ and sequentially built as

$$W_t = W_{t-1} \cdot B_t. \tag{23}$$

If the null hypothesis is true, the condition in Eq. 22 implies that no betting strategy should consistently increase wealth defined in Eq. 23. However, under an alternative $\mathbb{Q} \in \mathcal{Q}$, it is possible to increase wealth systematically. Therefore, **the wealth process $\{W_t\}_{t \in \mathcal{T}}$ can be interpreted as a measure of the evidence against the null**.

Following the testing-by-betting framework, we can construct an **empirically adaptive e-process** by combining several e-variables $\{E_t\}_{t \in \mathcal{T}}$ via the following martingale:

$$W_0 = 1, \quad W_t = \prod_{i=1}^{t}((1 - \lambda_i) + \lambda_i E_i). \tag{24}$$

Here the bet at each timestep t is $B_t = (1 - \lambda_t) + \lambda_t E_t$, where $\lambda_t \in [0, 1]$ (the betting rate) is a measurable function of $E_1, ..., E_{t-1}$.

To build this process, we need to choose a betting rate λ_t at each timestep. We do this by aiming for the optimal e-process that grows as fast as possible under the alternativeThis is called log-optimality in the e-values literature [33]. We can achieve this by maximizing the wealth (the bet evaluated at the realized value of W_t) at each timestep:

$$\lambda_t = \arg\max_{\lambda \in [0,1]} \mathbb{E}_{H_1}\left[\log W_t \mid \mathcal{F}_{t-1}\right] \tag{25}$$

Solving this optimization problem implies a significant computational expense, and therefore we usually opt for numerical approximations such as GRAPA [50] or SF-OGD [30], presented in § 3.3.

B. Derivation of the Empirically Adaptive e-process for Object Tracking Failure Detection

Our goal is to solve the hypothesis test in Eq. 1 by building an e-process $\{X_t\}_{t \geqslant 0}$.

We leverage the empirically adapted e-process framework presented in § A.2.4 to build an e-process for this test. For each timestep, we can define an e-value based on the discrepancy between the tracking metric at frame t and the threshold:

$$E_t = 1 + \varepsilon - m_t. \tag{26}$$

By the linearity of the expectation and Equation Eq. 1, E_t meets Definition A.6 and therefore is an e-value.

We can use the empirically adaptive e-process defined in Eq. 24 to merge these e-values:

$$\begin{aligned} W_t &= \prod_{i=1}^{t} ((1 - \lambda_i) + \lambda_i E_i) && (27) \\ &= \prod_{i=1}^{t} ((1 - \lambda_i) + \lambda_i (1 + \varepsilon - m_i)) && (28) \\ &= \prod_{i=1}^{t} (1 + \lambda_i (\varepsilon - m_i)), && (29) \end{aligned}$$

which results in the process defined in Eq. 5.