Detecting Rewards Deterioration in Episodic Reinforcement Learning

Ido Greenberg [1] **Shie Mannor** [1 2]

Abstract

In many RL applications, once training ends, it is vital to detect any deterioration in the agent performance as soon as possible. Furthermore, it often has to be done without modifying the policy and under minimal assumptions regarding the environment. In this paper, we address this problem by focusing directly on the rewards and testing for degradation. We consider an episodic framework, where the rewards within each episode are not independent, nor identically-distributed, nor Markov. We present this problem as a multivariate mean-shift detection problem with possibly partial observations. We define the mean-shift in a way corresponding to deterioration of a temporal signal (such as the rewards), and derive a test for this problem with optimal statistical power. Empirically, on deteriorated rewards in control problems (generated using various environment modifications), the test is demonstrated to be more powerful than standard tests – often by orders of magnitude. We also suggest a novel Bootstrap mechanism for False Alarm Rate control (BFAR), applicable to episodic (non-i.i.d) signal and allowing our test to run sequentially in an online manner. Our method does not rely on a learned model of the environment, is entirely external to the agent, and in fact can be applied to detect changes or drifts in any episodic signal.

1. Introduction

Reinforcement learning (RL) algorithms have recently demonstrated impressive success in a variety of sequential decision-making problems (Badia et al., 2020; Hessel et al., 2018). While most RL works focus on the maximization of rewards under various conditions, a key issue in real-world RL tasks is the safety and reliability of the system (Dulac-Arnold et al., 2019; Chan et al., 2020), arising in both offline and online settings.

In **offline settings**, comparing the agent performance in different environments is important for generalization (e.g., in sim-to-real and transfer learning). The comparison may indicate the difficulty of the problem or help to select the right learning algorithms. Uncertainty estimation, which could help to address this challenge, is currently considered a hard problem in RL, in particular for model-free methods (Yu et al., 2020).

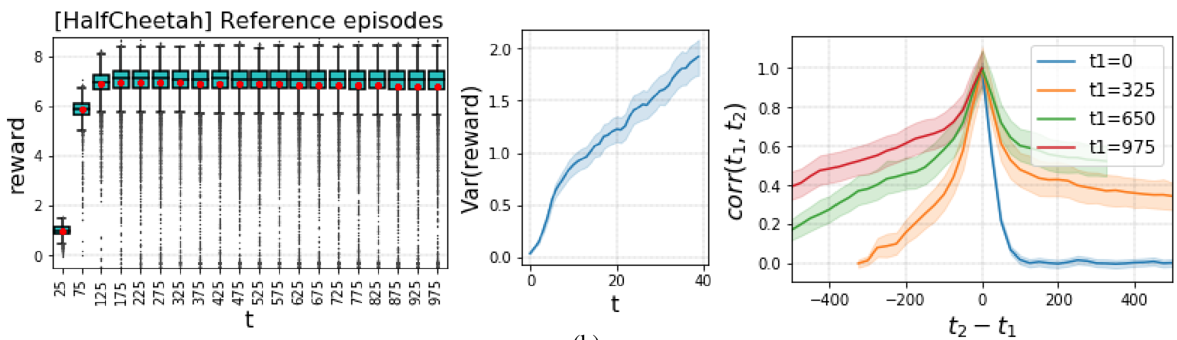
In **online settings**, where a fixed, already-trained agent runs continuously, its performance may be affected (gradually or abruptly) by changes in the controlled system or its surroundings, or when reaching unfamiliar states. Some works address robustness to changes (Lecarpentier & Rachelson, 2019; Lee et al., 2020), yet performance degradation is sometimes inevitable, and should be detected as soon as possible. The detection allows us to fall back into manual control, send the agent to re-train, guide diagnosis, or even bring the agent to halt. This problem is inherently different from robustness to changes during training: it focuses on safety and reliability, in post-training phase where intervention in the policy is limited or forbidden (Matsushima et al., 2020). It also operates in different time-scales: while training may take millions of episodes, changes should often be detected within tens of episodes, and critical failures – within less than an episode.

Such post-training performance-awareness is essential for any autonomous system in risk-intolerant applications, such as autonomous driving and medical devices. For example, when an autonomous car starts acting suspiciously with a passenger sitting inside, activating a training process and exploring for new policies is not an option. **The priority is to notice the suspicious behavior as soon as possible**, so that it can be alerted in time to save lives.

Many sequential statistical tests exist for detection of mean degradation in a random process. However, common methods (Page, 1954; Lan, 1994; Harel et al., 2014) assume independent and identically distributed (i.i.d) samples, while in RL the feedback from the environment is usually both highly correlated over consecutive time-steps, and varies over the life-time of the task (Korenkevych et al., 2019).

[1]Department of Electric Engineering, Technion, Israel [2]Nvidia Research. Correspondence to: Ido Greenberg <gido@campus.technion.ac.il>, Shie Mannor <shie@ee.technion.ac.il>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.



(a) (b) (c)

Figure 1: Properties of the rewards of a fixed agent in HalfCheetah, estimated over $N = 10000$ episodes of $T = 1000$ time-steps: (a) distribution of rewards per time-step; (b) variance per time-step; (c) correlation(t_1, t_2) vs. $t_2 - t_1$. The estimations are in resolution of 25 time-steps, i.e., every episode was split into 40 intervals of 25 consecutive steps, and each sample is the average over an interval.

This is demonstrated in Fig. 1.

A possible solution is to apply statistical tests to large blocks of data assumed to be i.i.d (Ditzler et al., 2015). This is particularly common in RL, where the episodic settings allow a natural blocks-partition (see for example Colas et al. (2019)). However, this approach requires complete episodes for change detection, while a faster response is often required. Furthermore, naively applying a statistical test on the accumulated feedback (e.g., sum of rewards) from complete episodes, ignores the dependencies within the episodes and misses vital information, leading to highly sub-optimal tests (as demonstrated in Section 6.2).

In this work, we devise an optimal test for detection of degradation of the rewards in an episodic RL task (or in any other episodic signal), based on the covariance structure within the episodes. Even in absence of the assumptions that guarantee its optimality, the test is still asymptotically superior to the common approach of comparing the mean reward (Colas et al., 2019). The test can detect changes and drifts in both the offline and the online settings defined above. Since tuning of the False Alarm Rate (FAR) of a sequential test usually relies on the underlying signal being i.i.d, we also suggest a novel Bootstrap mechanism for FAR control (BFAR) in sequential tests on episodic signals. The suggested procedures rely on the ability to estimate the correlations within the episodes, e.g., through a ”reference dataset” of episodes.

Since the test is applied directly to the rewards, it is model-free in the following senses: the underlying process is not assumed to be known, to be Markov, or to be observable at all (as opposed to other works, e.g., Banerjee et al. (2016)), and we require no knowledge about the process or the running policy. Furthermore, as the rewards are simply referred to as episodic time-series, the test can be similarly applied to detect changes in any episodic signal.

We demonstrate the new procedures in the environments of Pendulum (OpenAI), HalfCheetah and Humanoid (MuJoCo; Todorov et al., 2012). BFAR is shown to successfully control the false alarm rate. The suggested test detects degradation faster and more often than three alternative tests – in certain cases by orders of magnitude.

The paper is organized as follows: Section 3 formulates the offline setup (individual tests) and the online setup (sequential tests). Section 4 defines the model of an episodic signal, and derives an optimal degradation-test for such a signal. Section 5 shows how to adjust the test for online settings and control the false alarm rate. Section 6 describes the experiments, and Section 7 discusses related works.

To the best of our knowledge, we are the first to exploit the covariance between rewards in post-training phase to test for changes in RL-based systems. Our main contribution is an optimal test that can detect deterioration in agent rewards and other episodic signals reliably, in much shorter times than current standard tests. We also suggest a novel bootstrap mechanism to control false alarm rate of such tests on episodic (non-i.i.d) data. Finally, we lay a new framework for statistical tests on episodic signals, which opens the way for further research on this problem.

2. Preliminaries

Reinforcement learning and episodic framework: A Reinforcement Learning (RL) problem is usually modeled as a sequential *decision process*, where a learning *agent* has to repeatedly make decisions that affect its future states and rewards. The process is often organized as a finite sequence of time-steps (an *episode*) that repeats multiple times in different variants, e.g., with different initial states. Common examples are board and video games (Brockman et al., 2016), as well as more realistic problems such as autonomous driving tasks.

Once the agent is fixed (which is the case in this work), the rewards of the decision process essentially reduce to a (decision-free) random process $\{X_t\}_{t=1}^n$, which can be defined by its PDF ($f_{\{X_t\}_{t=1}^n} : \mathbb{R}^n \rightarrow [0, \infty)$). $\{X_t\}$ usually depend on each other: even in the popular *Markov Decision Process* (Bellman, 1957), where the dependence goes only a single step back, long-term correlations may still carry information if the states are not observable by the agent.

Hypothesis tests: Consider a parametric probability function $p(X|\theta)$ describing a random process, and consider two different hypotheses H_0, H_A determining the value (*simple hypothesis*) or allowed values (*complex hypothesis*) of θ. When designing a test to decide between the hypotheses, the basic metrics for the test efficacy are its *significance* $P(\text{not reject } H_0|H_0) = 1 - \alpha$ and its *power* $P(\text{reject } H_0|H_A) = \beta$. A hypothesis test with significance $1 - \alpha$ and power β is *optimal* if any test with as high significance $1 - \tilde{\alpha} \geq 1 - \alpha$ has smaller power $\tilde{\beta} \leq \beta$.

The likelihood of the hypothesis $H : \theta \in \Theta$ given data X is defined as $L(H|X) = \sup_{\theta\in\Theta} p(X|\theta)$. According to Neyman-Pearson lemma (Neyman et al., 1933), a threshold-test on the likelihood ratio $LR(H_0, H_A|X) = L(H_0|X)/L(H_A|X)$ is optimal. The threshold is uniquely determined by the desired significance level α, though is often difficult to practically calculate given α.

In many practical applications, a hypothesis test is repeatedly applied as the data change or grow, a procedure known as a *sequential test*. If the null hypothesis H_0 is true, and any individual hypothesis test falsely rejects H_0 with some probability α, then the probability that at least one of the multiple tests will reject H_0 is $\alpha_0 > \alpha$, termed *family-wise type-I error* (or *false alarm rate* when associated with frequency). See Appendix A for more details about hypothesis testing and sequential tests in particular.

Common approaches for sequential tests, such as CUSUM (Page, 1954; Ryan, 2011) and α-spending functions (Lan, 1994; Pocock, 1977), usually require strong assumptions such as independence or normality, as further discussed in Appendix B.

3. Problem Setup

In this work, we consider two setups where detecting performance deterioration is important – sequential degradation-tests and individual degradation-tests. The individual tests, in addition to their importance in offline settings such as sim-to-real and transfer learning, are used in this work as building-blocks for the online sequential tests.

Both setups assume a fixed agent that was previously trained, and aim to detect whenever the agent performance begins to deteriorate, e.g., due to environment changes. The ability to notice such changes is essential in many real-world problems, as explained in Section 1.

Setup 1 (**Individual degradation-test).** We consider a fixed trained agent (policy must be fixed but is not necessarily optimal), whose rewards in an episodic environment (with episodes of length T) were previously recorded for multiple episodes (the *reference dataset*). The agent runs in a new environment for n time-steps (both $n < T$ and $n \geq T$ are valid). The goal is to decide whether the rewards in the new environment are smaller than the original environment or not. If the new environment is identical, the probability of a false alarm must not exceed α.

Setup 2 (**Sequential degradation-test).** As in Setup 1, we consider a fixed trained agent with reference data of multiple episodes. This time the agent keeps running in the same environment, and at a certain point in time its rewards begin to deteriorate, e.g., due to changes in the environment. The goal is to alert to the degradation as soon as possible. As long as the environment has not changed, the probability of a false alarm must not exceed α_0 per $\tilde{h}$ episodes.

Note that while in this work the setups focus on degradation, they can be easily modified to look for any change (as positive changes may also indicate the need for further training, for example).

4. Optimization of Individual Tests

To tackle the problem of Setup 1, we first define the properties of an episodic signal and the general assumptions regarding its degradation.

Definition 4.1 (T-long episodic signal)**.** *Let $n, T \in \mathbb{N}$, and write $n = KT + \tau_0$ (for non-negative integers K, τ_0 with $\tau_0 \leq T$). A sequence of real-valued random variables $\{X_t\}_{t=1}^n$ is a T-long episodic signal, if its joint probability density function can be written as*

$$f_{\{X_t\}_{t=1}^n}(x_1, ..., x_n) = \left[\prod_{k=0}^{K-1} f_{\{X_t\}_{t=1}^T}(\{x_{kT+t}\}_{t=1}^T)\right] \cdot f_{\{X_t\}_{t=1}^{\tau_0}}(\{x_{KT+t}\}_{t=1}^{\tau_0}) \tag{1}$$

(where an empty product is defined as 1). We further denote $\boldsymbol{\mu_0} := E[(X_1, ..., X_T)^\top] \in \mathbb{R}^T$, $\Sigma_0 := Cov((X_1, ..., X_T)^\top, (X_1, ..., X_T)) \in \mathbb{R}^{T\times T}$.

Note that the episodic signal consists of i.i.d episodes, but is not assumed to be independent or identically-distributed within the episodes – a setup particularly popular in RL.

In the analysis below we assume that both $\boldsymbol{\mu_0}$ and Σ_0 are known. This can be achieved either with detailed domain knowledge, or by estimation from the recorded reference dataset of Setup 1, assuming it satisfies Eq. (1). The estimation errors decrease as $\mathcal{O}(1/\sqrt{N})$ with the number N of reference episodes, and are distributed according to the Central Limit Theorem (for means) and Wishart distribution (K. V. Mardia & Bibby, 1979) (for covariance). While in this work we use up to $N = 10000$ reference episodes, Appendix J shows that $N = 300$ reference episodes are sufficient for reasonable results in HalfCheetah, for example. Note that correlations estimation has been already discussed in several other RL works (Alt et al., 2019).

Fig. 1 demonstrates the estimation of mean and covariance parameters for a trained agent in the environment of HalfCheetah, from a reference dataset of $N = 10000$ episodes. This also demonstrates the non-trivial correlations structure in the environment. According to Fig. 1b, the variance in the rewards varies and does not seem to reach stationarity within the scope of an episode. Fig. 1c shows the autocorrelation function $ACF(t_2 - t_1) = corr(t_1, t_2)$ for different reference times t_1. The correlations clearly last for hundreds of time-steps, and depend on the time t_1 rather than merely on the time-difference $t_2 - t_1$. This means that the autocorrelation function is not expressive enough for the actual correlations structure.

Once the per-episode parameters $\boldsymbol{\mu_0} \in \mathbb{R}^T, \Sigma_0 \in \mathbb{R}^{T\times T}$ are known, the mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n\times n}$ of the whole signal can be derived directly: $\boldsymbol{\mu}$ consists of periodic repetitions of $\boldsymbol{\mu_0}$, and Σ consists of copies of Σ_0 as $T\times T$ blocks along its diagonal. For both, the last repetition is cropped if n is not an integer multiplication of T. In other words, by taking advantage of the episodic setup, we can treat the temporal univariate non-i.i.d signal as a multivariate signal with easily-measured mean and covariance – even if the signal ends in the middle of an episode.

The degradation in the signal $X = \{X_t\}_{t=1}^n$ is defined through the difference between two hypotheses. The null hypothesis H_0 states that X is a T-long episodic signal with expectations $\boldsymbol{\mu_0} \in R^T$ and invertible covariance matrix $\Sigma_0 \in R^{T\times T}$. Our first alternative hypothesis (H_A) – uniform degradation – states that X is a T-long episodic signal with the same covariance Σ_0 but smaller expectations: $\exists\epsilon \geq \epsilon_0, \forall 1 \leq t \leq T : (\tilde{\boldsymbol{\mu}}_{\mathbf{0}})_t = (\boldsymbol{\mu_0})_t - \epsilon$. Note that this hypothesis is complex ($\epsilon \geq \epsilon_0$), where ϵ_0 can be tuned according to the minimal degradation magnitude of interest. In fact, Theorem 4.1 shows that the optimal corresponding test is independent of the choice of ϵ_0.

Theorem 4.1 (Optimal test for uniform degradation)**.** *Define the uniform-degradation weighted-mean $s_{unif}(X) := W \cdot X$, where $W := \mathbf{1}^\top \cdot \Sigma^{-1} \in \mathbb{R}^n$ (and $\mathbf{1}$ is the all-1 vector). If the distribution of X is multivariate normal, then a threshold-test on s_{unif} is optimal.*

Proof Sketch (see full proof in Appendix E). According to Neyman-Pearson lemma (Neyman et al., 1933), a threshold-test on the likelihood-ratio (LR) between H_0 and H_A is optimal. Since H_A is complex, the LR is a minimum over $\epsilon \in [\epsilon_0, \infty)$. Lemma 1 shows that $\exists s_0 : s_{unif} \geq s_0 \Rightarrow \epsilon = \epsilon_0$ and $s_{unif} \leq s_0 \Rightarrow \epsilon = \epsilon(s_{unif})$. The rest of the proof substitutes ϵ in both domains of s_{unif} to prove monotony of the LR in s_{unif}, from which we can conclude monotony in s_{unif} over all $\mathbb{R}$. □

Following Theorem 4.1, we define the Uniform Degradation Test (**UDT**) to be a threshold-test on s_{unif}, i.e., "declare a degradation if $s_{unif} < \kappa$" for a pre-defined κ. If the weights are calculated in advance, s_{unif} can be calculated in $\mathcal{O}(n)$ time, and updated in $\mathcal{O}(1)$ with every new sample.

Recall that test optimality is defined in Section 2 as having maximal power per significance level. To achieve the significance α required in Setup 1, we apply a bootstrap mechanism that randomly samples episodes from the reference data and calculates the corresponding statistic (e.g., s_{unif}). This yields a bootstrap-estimate of the statistic's distribution under H_0, and the α-quantile of the estimated distribution is chosen as the test-threshold ($\kappa = q_\alpha(s_{unif}|H_0)$).

H_A is intended for degradation in a temporal signal, and derives a different optimal statistic than standard mean-change tests in multivariate variables (e.g., Hotelling). In Section 6, this is indeed demonstrated to be more powerful for rewards degradation. Also note that by explicitly referring to the temporal dimension, we allow detections even before the first episode is completed.

Theorem 4.1 relies on multivariate normality assumption, which is often too strong for real-world applications. Theorem 4.2 guarantees that if we remove the normality assumption, it is still beneficial to look into the episodes instead of considering them as atomic blocks; that is, UDT is still asymptotically better than a test on the simple mean $s_{simp} = \sum_{t=1}^n X_t/n$. Note that "asymptotic" refers to the signal length $n \to \infty$ (while T remains constant), and is translated in the sequential setup into a "very long lookback-horizon h" (rather than very long running time).

Theorem 4.2 (Asymptotic power of UDT)**.** *Denote the length of the signal $n = K \cdot T$, assume a uniform degradation of size $\frac{\epsilon}{\sqrt{K}}$, and let two threshold-tests τ_{simp} on s_{simp} and UDT on s_{unif} be tuned to have significance α. Then*

$$\lim_{K\to\infty} P\left(\tau_{simp} \text{ rejects } \middle| H_A\right) = \Phi\left(q_\alpha^0 + \frac{\epsilon T}{\sqrt{\mathbf{1}^\top \Sigma_0 \mathbf{1}}}\right)$$
$$\leq \Phi\left(q_\alpha^0 + \epsilon\sqrt{\mathbf{1}^\top \Sigma_0^{-1} \mathbf{1}}\right) = \lim_{K\to\infty} P\left(\text{UDT rejects } \middle| H_A\right) \tag{2}$$

where Φ is the CDF of the standard normal distribution, and q_α^0 is its α-quantile.

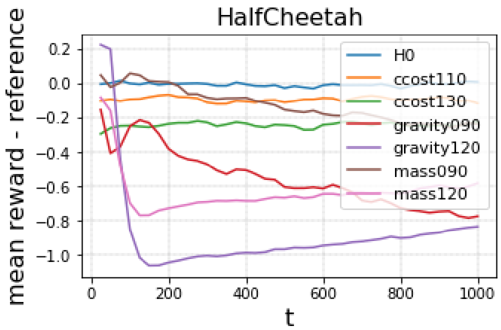


Figure 2: Rewards degradation of a fixed agent in HalfCheetah following changes in gravity, mass, and control-cost, over $N = 5000$ episodes per scenario.

Proof Sketch (see full proof in Appendix E). Since the episodes of the signal are i.i.d, both s_{simp} and s_{unif} are asymptotically normal according to the Central Limit Theorem. The means and variances of both statistics are calculated in Lemma 2. Calculation of the variance of s_{unif} relies on writing s_{unif} as a sum of linear transformations of X ($s_{unif} = \sum_{i=1}^{n}(\Sigma^{-1})_i X$), and using the relation between Σ and Σ_0. The inequality between the resulted powers is shown to be equivalent to a matrix-form of the means-inequality, and is proved using Cauchy-Schwarz inequality for $\Sigma_0^{-1/2}\mathbf{1}$ and $\Sigma_0^{1/2}\mathbf{1}$. $\square$

Motivated by Theorem 4.2, we define $G^2 := \frac{(\mathbf{1}^\top \Sigma_0^{-1}\mathbf{1})(\mathbf{1}^\top \Sigma_0 \mathbf{1})}{T^2}$ to be the asymptotic power gain of UDT, quantify it, and show that it increases with the heterogeneity of the spectrum of Σ_0. In particular, if the rewards are heterogeneous, the suggested test is guaranteed to detect uniform degradation with much higher probability than the standard mean-test.

Proposition 4.1 (Asymptotic power gain)**.** *$G^2 = 1 + \sum_{i,j=1}^{T} w_{ij}(\lambda_i - \lambda_j)^2$, where $\{\lambda_i\}_{i=1}^{T}$ are the eigenvalues of Σ_0 and $\{w_{ij}\}_{i,j=1}^{T}$ are positive weights.*

Proof Sketch (see full proof in Appendix E). The result can be calculated after diagonalization of Σ_0, and the weights $\{w_{ij}\}$ are derived from the diagonalizing matrix. $\square$

So far we assumed uniform degradation. In the context of RL, such a model may refer to changes in constant costs or action costs, as well as certain dynamics whose change influences various states in a similar way. Fig. 2 demonstrates the empiric degradation in the rewards of a fixed agent in HalfCheetah, following changes in gravity, mass and control-cost. It seems that some modifications indeed cause a quite uniform degradation, while in others the degradation is mostly restricted to certain ranges of time.

To model effects that are less uniform in time we suggest a partial degradation hypothesis, where some (unknown) entries of $\boldsymbol{\mu_0}$ are reduced by $\epsilon > 0$, and others do not change. The number $m = p \cdot T$ of the reduced entries is defined by a parameter $p \in (0, 1)$.

This time, calculation of the optimal test-statistic through the LR yields a minimum over $\binom{T}{m}$ possible subsets of decreased entries, which is computationally heavy. However, Theorem 4.3 shows that if we optimize for small values of ϵ (where optimality is indeed most valuable), a near-optimal statistic is s_{part}, which is the sum of the $m = p \cdot T$ smallest time-steps of $(X - \boldsymbol{\mu})$ after a Σ^{-1}-transformation (see formal definition in Definition D.11). The resulted time-complexity is $\mathcal{O}(nT)$. We define the Partial Degradation Test (**PDT**) as a threshold-test on s_{part} with a parameter p.

Theorem 4.3 (Near-optimal test for uniform degradation)**.** *Assume that X is multivariate normal, and let P_α be the maximal power of a hypothesis test with significance $1-\alpha$. The power of a threshold-test on s_{part} with significance $1 - \alpha$ is $P_\alpha - \mathcal{O}(\epsilon)$.*

Proof Sketch. The expression to be minimized is shown to be the sum of two terms. One term is the sum of a subset of entries of $\Sigma^{-1}(X - \boldsymbol{\mu})$, which is minimized by simply taking the lowest entries (up to the constraint of consistency across episodes, which requires us to sum the rewards per time-step in advance). In Appendix E we bound the second term and its effects on the modified statistic and on the modified test-threshold. We show that the resulted decrease of rejection probability is $\mathcal{O}(\epsilon)$. $\square$

5. Bootstrap for False Alarm Rate Control (BFAR)

For Setup 2, we suggest a sequential testing procedure: run an individual test every d steps (i.e., $F = T/d$ test-points per episode), and return once any individual test declares a degradation. The tests can run according to Section 4, applied on the h recent episodes. Multiple tests may be applied every test-point, e.g., with varying test-statistics $\{s\}$ or lookback-horizons $\{h\}$. This procedure, as implemented for the experiments of Section 6, is described in Fig. 3.

Setup 2 limits the probability of a false alarm to α_0 in a run of $\tilde{h}$ episodes. To satisfy this condition, we set a uniform threshold κ on the p-values of the individual tests (i.e., declare once a test returns p-val $< \kappa$). The threshold is determined using a Bootstrap mechanism for False Alarm control (**BFAR**, Algorithm 1).

While bootstrap methods for false alarm control are quite popular, they often rely on the data samples being i.i.d (Kharitonov et al., 2015; Abhishek & Mannor, 2017), which is crucial for the re-sampling to reliably mimic the source of the signal. To address the non-i.i.d signal, we take advantage of the episodic framework and sample whole episodes. We then use the re-sampled sequence to simulate tests on sub-sequences where the first and last episodes may be incomplete, as described below. This allows simulation of sequences of various lengths (including non-integer number of episodes) without assuming independence, normality, or identical distributions within the episodes.

Algorithm 1: BFAR: Bootstrap for FAR control

Input: reference dataset $x \in \mathbb{R}^{N \times T}$; statistic functions $\{s\}$; lookback-horizons $\{h_1, ..., h_{max}\}$; test length $\tilde{h} \in \mathbb{N}$; bootstrap repetitions $B \in \mathbb{N}$; desired significance $\alpha_0 \in (0, 1)$;
Output: test threshold for individual tests;
Initialize $P = (1, ..., 1) \in [0, 1]^B$;
for *b in 1:B* **do**
    Initialize $Y \in \mathbb{R}^{(h_{max}+\tilde{h})T}$;
    for *k in 0:(h_{max}+$\tilde{h}$-1)* **do**
        Sample j uniformly from $(1, ..., N)$;
        $Y[kT + 1 : kT + T] \leftarrow (x_{j1}, ..., x_{jT})$;
    for *t in test-points* **do**
        for *h in lookback-horizons and s in statistic functions* **do**
            $y \leftarrow Y[t - hT : t]$;
            $p \leftarrow$ individual_test_pvalue(y vs. $x; s$)
            $P[b] \leftarrow \min(P[b], p)$;
Return $quantile_{\alpha_0}(P)$;

BFAR samples $h_{max} + \tilde{h}$ episodes (where h_{max} is the maximal lookback-horizon) from reference data of N episodes, to simulate sequential data Y. Then individual tests are simulated for any test-point along $\tilde{h}$ episodes, starting after h_{max} episodes. The minimal p-value determines whether a detection would occur in Y. The whole procedure repeats B times, creating a bootstrap estimate of the distribution of the minimal p-value along $\tilde{h}$ episodes. We choose the tests threshold to be the α_0-quantile of this distribution, such that α_0 of the bootstrap simulations would raise a false alarm.

Note that the statistic for the tests is given to BFAR as an input, making its choice independent of BFAR. BFAR can run in an offline manner (e.g., a single run before the deployment of the agent). It takes $\mathcal{O}(BF\tilde{h}\tilde{T})$ time, where $\tilde{T}$ is the time of a single update of all the test-statistics. Additional details are discussed in Appendices F,G.

6. Experiments

6.1. Methodology

We run experiments in standard RL environments as described below. For each environment, we train an agent using the PyTorch version (Kostrikov, 2018) of OpenAI's baseline (Dhariwal et al., 2017) of A2C algorithm (Mnih et al., 2016). We let the trained agent run in the environment for N_0 episodes and record its rewards, considered the *trusted reference data*. We then define several scenarios, and let the agent run for $M \times N$ episodes in each scenario (divided later into $M = 100$ blocks of N episodes). One scenario is named H_0 and is identical to the reference up to the random initial-states. The other scenarios are defined per environment, and present environmental changes expected to harm the agent's rewards. The agent is *not* trained to adapt to these changes, and the goal is to test how long it takes for a degradation-test to detect the degradation.

Individual degradation-tests of length n (Setup 1) are applied for every scenario over the first n time-steps of each block. Sequential degradation-tests (Setup 2) are applied sequentially over the episodes of each block. Since the agent is assumed to run continuously as the environment changes from H_0 to an alternative scenario, each block is preceded by a random sample of H_0 episodes, as demonstrated in Fig. 3.

BFAR adjusts the tests thresholds to have a false alarm with probability $\alpha_0 = 5\%$ per $\tilde{h} = N$ episodes (where N is the data-block size). Two lookback-horizons h_1, h_2 are chosen for every environment. The rewards are downsampled by a factor of d before applying the tests, intended to reduce the parameters estimation error. Table 1 summarizes the setup of the various environments.

The experimented degradation-tests are a threshold-test on the simple **Mean**; **CUSUM** (Ryan, 2011); **Hotelling** (Hotelling, 1931); **UDT** and **PDT** (with $p = 0.9$) from Section 4; and a Mixed Degradation Test (**MDT**) that runs Mean, Hotelling and PDT in parallel – applying all three in every test-point (as permitted in Algorithm 1). All the degradation-tests are tuned according to the same reference data. Further implementation details are discussed in Appendix H.

6.2. Results

We run the tests in the environments of Pendulum (OpenAI), where the goal is to keep a pendulum pointing upwards; HalfCheetah (Todorov et al., 2012), where the goal

Table 1: Environments parameters
(episode length (T), reference episodes (N_0), test blocks (M), episodes per block (N), sequential test length ($\tilde{h}$), lookback horizons (h_1, h_2), tests per episode ($F = T/d$))

Environment	T	N_0	M	$N = \tilde{h}$	$h_{1,2}$	F
Pendulum	200	3e3	100	30	3,30	20
HalfCheetah	1000	1e4	100	50	5,50	40
Humanoid	200	5e3	100	30	3,30	10

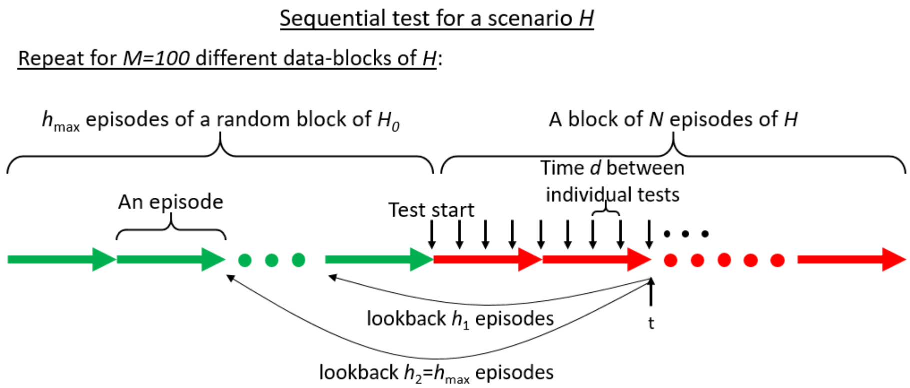


Figure 3: A summary of the sequential degradation-test procedure described in Section 6.1.

is for a 2D cheetah to run as fast as possible; and Humanoid, where the goal is for a person to walk without falling. In each environment we define the scenario *ccostx* of control cost increased to x% of its original value, as well as changed-dynamics scenarios specified in Appendix H.

In all the environments the rewards are clearly *not* independent, identically-distributed or normal (see Fig. 1 for example). Yet the false alarm rates are close to $\alpha_0 = 5\%$ per $\tilde{h}$ episodes in all the tests, as demonstrated in Fig. 4 (and in more details in Fig. 6 in Appendix I). These results under H_0 indicate that **BFAR tunes the thresholds properly** in spite of the complexity of the data. Note that BFAR never observed the data of scenario H_0 – only the reference data.

In most of the non-H_0 scenarios, our tests prove to be more powerful than the standard tests, often by extreme margins. For example, increased control cost in all the environments and additive noise in Pendulum are all 100%-detected by the suggested tests, usually within few episodes (Fig. 4); whereas Mean, CUSUM and Hotelling have very poor detection rates. Mean did not detect degradation in Pendulum even after the control cost increased from 110% to 300%(!), while keeping the significance level constant ($\alpha_0 = 5\%$).

Note that we run the tests with two lookback-horizons in parallel, as allowed by BFAR. This proves useful: with +30% control cost in HalfCheetah, for example, the short lookback-horizon allows fast detection of degradation; but with merely +10%, the long horizon is necessary to notice the slight degradation over a large number of episodes. This is demonstrated in Fig. 11 in Appendix I.

Covariance-based tests reduce the weights of the highly-varying (and presumably noisier) time-steps. In HalfCheetah they turn out to be in the later parts of the episode. As a result, in certain scenarios, Mean, CUSUM and Hotelling (which do not exploit the different variances optimally) do better in individual tests of 100 samples (out of $T = 1000$) than they do in one or even 10 full episodes (see Fig. 10a in Appendix I). This does not occur in UDT and PDT. Es-

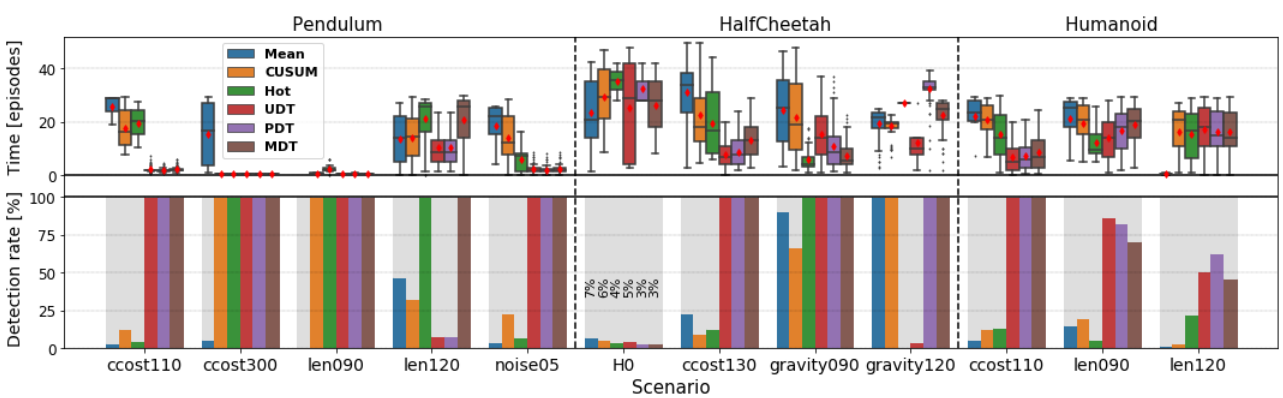


Figure 4: **Bottom**: percent of sequential tests that ended with degradation detection (high is good), over $M = 100$ runs with different seeds, for 3 standard tests and 3 variants of our test (**UDT**, **PDT** and **MDT**), in a sample of scenarios in Pendulum, HalfCheetah and Humanoid. **Top**: time until detection (low is good) – for the runs that ended with detection. The significance of the tests is shown for HalfCheetah in H_0 scenario (and for Pendulum and Humanoid as well in Fig. 6 in Appendix I).

sentially, we see that **ignoring the noise variability leads to violation of the principle that more data are better**.

In Pendulum, the ratio between variance of different steps may reach 5 orders of magnitude. This phenomenon increases the potential power of the covariance-based tests. For example, when the pole is shortened, negative changes in the highly-weighted time-steps are detected even when the mean of the whole signal increases. This feature allows us to detect slight changes in the environment before they develop into larger changes and cause damage.

On the other hand, a challenging situation arises when certain rewards decrease but the highly-weighted ones slightly increase (as in longer Pendulum's pole), which strongly violates the assumptions of Section 4. UDT is doomed to falter in such scenarios. PDT proves somewhat robust to this phenomenon since it is capable of focusing on a subset of time-steps, as demonstrated in increased gravity in HalfCheetah (Fig. 4). However, it cannot overcome the extreme weights differences in Pendulum. The one test that demonstrated robustness to all the experimented scenarios, including modified Pendulum's length and mass, is MDT. MDT combines Mean, Hotelling and PDT and does not fall far behind any of the three, in any of the scenarios. Hence, it presents excellent results in some scenarios and reasonable results in the others.

The tests were run on a single i9-10900X CPU core. BFAR (which needs to run only once and in an offline manner – before the deployment of the agent) took around 30 minutes per environment and test-statistic (several hours in total). Any parallelization should accelerate the bootstrap linearly with the number of cores. The sequential (online) tests themselves ran for 10 minutes per scenario – for all the 6 test-statistics together and for thousands of episodes.

Detailed experiments results are available in Appendix I. The code of the experiments is available on GitHub.

7. Related Work

Training in non-stationary environments has been widely researched, in particular in the frameworks of Multi-Armed Bandits (Mukherjee & Maillard, 2019; Garivier & Moulines, 2011; Besbes et al., 2014; Lykouris et al., 2020; Alatur et al., 2020; Gupta et al., 2019; Jun et al., 2018), model-based RL (Lecarpentier & Rachelson, 2019; Lee et al., 2020) and general multi-agent environments (Hernandez-Leal et al., 2019). Banerjee et al. (2016) explicitly detect changes in the environment and modify the policy accordingly, but assume that the environment is Markov, fully-observable, and its transition model is known – three assumptions that we avoid and that do not hold in many real-world problems. Safe exploration during training in RL was addressed by Garcia & Fernandez (2015); Chow et al. (2018); Junges et al. (2016); Cheng et al. (2019); Alshiekh (2017). Note that our work refers to changes beyond the scope of the training phase: it addresses the stage where the agent is fixed and required not to train further, in particular not in an online manner. Robust algorithms may prevent degradation in the first place, but when they fail – or when their assumptions are not met – an external model-free monitor with minimal assumptions (as the one suggested in this work) is crucial.

Sequential tests were addressed by many over the years. Common approaches rely on strong assumptions such as samples independence (Page, 1954; Ryan, 2011) and normality (Pocock, 1977; O'Brien & Fleming, 1979). Generalizations exist for certain private cases (Lu & Jr., 2001; Xie & Siegmund, 2011), sometimes at cost of alternative assumptions such as known change-size (Lund et al., 2007). Samples independence is usually assumed also in recent works with numeric approaches (Kharitonov et al., 2015; Abhishek & Mannor, 2017; Harel et al., 2014), and is often justified by consolidating many samples (e.g., an episode) together as a single sample (Colas et al., 2019). Ditzler et al. (2015) wrote that "change detection is typically carried out by inspecting i.i.d features extracted from the incoming data stream, e.g., the sample mean". Certain works address cyclic signals monitoring (Zhou et al., 2005), but to the best of our knowledge, we are the first to devise an optimal test for mean change in temporal non-i.i.d signals, and a false alarm control mechanism for such non-i.i.d signals.

Our work can be seen in part as converting a univariate temporal episodic signal into a T-dimensional multivariate signal. Many works addressed the problem of **change-point detection in multivariate variables**, e.g., using histograms comparison (Boracchi et al., 2018), Hotelling statistic (Hotelling, 1931), and K-L distance (Kuncheva, 2013). Hotelling in particular also looks for changed mean under unchanged covariance. However, unlike existing tests, we derive optimal tests for two different negative mean-change hypotheses, intended to detect degradation in temporal signals. Indeed, Section 6 demonstrates the advantage over Hotelling in such a context. In addition, by considering the temporal nature of the signal, we are able to handle "incomplete observations" and in particular obtain detections even within the middle of the first episode.

8. Summary

We introduced a novel approach that is optimal (under certain conditions) for detection of changes in episodic signals, exploiting the correlations structure as measured in a reference dataset. In environments of classic control (Pendulum) and MuJoCo (HalfCheetah, Humanoid), the suggested statistical tests detected degradation faster than alternatives, often by orders of magnitude. Certain conditions, such as combination of positive and negative changes in very heterogeneous signals, may cause instability in some of the suggested tests; however, this is shown to be solved by running the new test in parallel to standard tests – with only a small loss of test power.

We also introduced BFAR, a bootstrap mechanism that adjusts tests thresholds according to the desired false alarm rate in sequential tests. The mechanism empirically succeeded in providing valid thresholds for various tests in all the environments, in spite of the non-i.i.d data.

The suggested approach may contribute to development of reliable RL-based systems. Future research may consider different hypotheses, such as a permitted small degradation (instead of H_0) or a mix of degradation and improvement (instead of H_A); suggest additional stabilizing mechanisms for covariance-based tests; exploit other metrics than rewards for tests on model-based RL systems; and apply comparative tests of episodic signals beyond the scope of sequential change detection.

Acknowledgements

This work was partially funded by the Israel Science Foundation under ISF grant number 2199/20. The authors wish to thank Guy Tennenholtz and Nadav Merlis for their helpful insights.

References

Abhishek, V. and Mannor, S. A nonparametric sequential test for online randomized experiments. *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 610–6, 2017.

Alatur, P., Levy, K. Y., and Krause, A. Multi-player bandits: The adversarial case. *JMLR*, 2020.

Alshiekh, M. Safe reinforcement learning via shielding. *Logic in Computer Science*, 2017.

Alt, B., Sosic, A., and Koeppl, H. Correlation priors for reinforcement learning. *NeurIPS*, 2019.

Aminikhanghahi, S. and Cook, D. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51:339–367, 2016.

Badia, A. P. et al. Agent57: Outperforming the atari human benchmark. *ICML*, 2020.

Banerjee, T., Liu, M., and How, J. Quickest change detection approach to optimal control in markov decision processes with model changes, 09 2016.

Bellman, R. A markovian decision process. *Indiana Univ. Math. J.*, 6:679–684, 1957. ISSN 0022-2518.

Berry, D. A. and Fristedt, B. *Bandit problems*. Springer Netherlands, 1985. doi: 10.1007/978-94-015-3711-7.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems (NIPS)*, 27, 2014.

Boracchi, G., Carrera, D., Cervellera, C., and Maccio, D. Quanttree: Histograms for change detection in multivariate data streams. *Proceedings of Machine Learning Research*, 80:639–648, 10–15 Jul 2018. URL `http://proceedings.mlr.press/v80/boracchi18a.html`.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym, 2016.

Brook, D. et al. An approach to the probability distribution of cusum run length. *Biometrika*, 59(3):539–549, 1972.

Bylander, T. Lecture notes: Reinforcement learning. `http://www.cs.utsa.edu/~bylander/cs6243/reinforcement-learning.pdf`.

Chan, S. C. et al. Measuring the reliability of reinforcement learning algorithms. *ICLR*, 2020.

Chen, J. Conditional value at risk (cvar). `https://www.investopedia.com/terms/c/conditional_value_at_risk.asp`, 2020.

Cheng, R. et al. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. *AAAI Conference on Artificial Intelligence*, 2019.

Chow, Y. et al. A lyapunov-based approach to safe reinforcement learning. *NIPS*, 2018.

Colas, C., Sigaud, O., and Oudeyer, P.-Y. A hitchhiker's guide to statistical comparisons of reinforcement learning algorithms, 2019.

Dai, B., Ding, S., and Wahba, G. Multivariate bernoulli distribution. *Bernoulli*, 19(4):1465–1483, 09 2013. doi: 10.3150/12-BEJSP10. URL `https://doi.org/10.3150/12-BEJSP10`.

Dhariwal, P., Hesse, C., Klimov, O., Nichol, A., Plappert, M., Radford, A., Schulman, J., Sidor, S., Wu, Y., and Zhokhov, P. Openai baselines. `https://github.com/openai/baselines`, 2017.

Dickey, D. A. and Fuller, W. A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431, 1979. doi: 10.1080/01621459.1979.

10482531. URL https://doi.org/10.1080/01621459.1979.10482531.

Ditzler, G., Polikar, R., and Alippi, C. Learning in nonstationary environments: A survey. *IEEE Computational Intelligence Magazine*, 2015.

Dulac-Arnold, G., Mankowitz, D., and Hester, T. Challenges of real-world reinforcement learning, 2019.

Efron, B. Second thoughts on the bootstrap. *Statist. Sci.*, 18(2):135–140, 05 2003. doi: 10.1214/ss/1063994968. URL https://doi.org/10.1214/ss/1063994968.

Freedman, A. Convergence theorem for finite markov chains, 2017. URL https://math.uchicago.edu/~may/REU2017/REUPapers/Freedman.pdf.

Garcia, J. and Fernandez, F. A comprehensive survey on safe reinforcement learning. *JMLR*, 2015.

Garivier, A. and Moulines, E. On upper-confidence bound policies for switching bandit problems. *International Conference on Algorithmic Learning Theory*, pp. 174–188, 10 2011. doi: 10.1007/978-3-642-24412-4_16.

Goldman, M. Lecture notes in stat c141: The bonferroni correction. https://www.stat.berkeley.edu/ mgoldman/Section0402.pdf, 2008.

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. *Proceedings of Machine Learning Research*, 2019.

Harel, M., Crammer, K., El-Yaniv, R., and Mannor, S. Concept drift detection through resampling. *International Conference on Machine Learning*, pp. II–1009–II–1017, 2014.

Henderson, P. et al. Deep reinforcement learning that matters. *AAAI*, 2017.

Hernandez-Leal, P., Kaisers, M., Baarslag, T., and de Cote, E. M. A survey of learning in multiagent environments: Dealing with non-stationarity, 2019.

Hessel, M., Modayil, J., van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., and Silver, D. Rainbow: Combining improvements in deep reinforcement learning. *AAAI*, 2018.

Hotelling, H. The generalization of student's ratio. *Ann. Math. Statist.*, 2(3):360–378, 08 1931. doi: 10.1214/aoms/1177732979. URL https://doi.org/10.1214/aoms/1177732979.

Irwin, M. E. Lecture notes: Convergence in distribution and central limit theorem. http://www2.stat.duke.edu/~sayan/230/2017/Section53.pdf, 2006.

Jun, K.-S. et al. Adversarial attacks on stochastic bandits. *NeurIPS*, 2018.

Junges, S. et al. Safety-constrained reinforcement learning for mdps. *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, 2016.

K. V. Mardia, J. T. K. and Bibby, J. M. *Multivariate analysis*. Academic Press, 1979.

Kharitonov, E., Vorobev, A., Macdonald, C., Serdyukov, P., and Ounis, I. Sequential testing for early stopping of online experiments. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 473–482, 2015. doi: 10.1145/2766462.2767729. URL https://doi.org/10.1145/2766462.2767729.

Korenkevych, D., Mahmood, A. R., Vasan, G., and Bergstra, J. Autoregressive policies for continuous control deep reinforcement learning, 2019.

Kostrikov, I. Pytorch implementations of reinforcement learning algorithms. https://github.com/ikostrikov/pytorch-a2c-ppo-acktr-gail, 2018.

Kroese, D. P., Brereton, T., Taimre, T., and Botev, Z. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6: 386–392, 2014.

Kuncheva, L. I. Change detection in streaming multivariate data using likelihood detectors. *IEEE Transactions on Knowledge and Data Engineering*, 25(5):1175–1180, 2013. doi: 10.1109/TKDE.2011.226.

Lan, D. L. D. K. K. G. Interim analysis: The alpha spending function approach. *Statistics in Medicine*, 13:1341–52, 1994.

Lecarpentier, E. and Rachelson, E. Non-stationary markov decision processes: a worst-case approach using model-based reinforcement learning. *NeurIPS 2019*, abs/1904.10090, 2019. URL http://arxiv.org/abs/1904.10090.

Lee, K. et al. Context-aware dynamics model for generalization in model-based rl. *ICML*, 2020.

Lu, C.-W. and Jr., M. R. R. Cusum charts for monitoring an autocorrelated process. *Journal of Quality Technology*, 33(3):316–334, 2001. doi: 10.1080/00224065.2001.11980082. URL https://doi.org/10.1080/00224065.2001.11980082.

Lund, R., Wang, X. L., Lu, Q. Q., Reeves, J., Gallagher, C., and Feng, Y. Changepoint Detection in Periodic and Autocorrelated Time Series. *Journal of Climate*, 20(20): 5178–5190, 10 2007. ISSN 0894-8755. doi: 10.1175/JCLI4291.1. URL https://doi.org/10.1175/JCLI4291.1.

Lykouris, T., Mirrokni, V., and Leme, R. P. Bandits with adversarial scaling. *ICML*, 2020.

MathWorks. Conditional value-at-risk (cvar). https://www.mathworks.com/discovery/conditional-value-at-risk.html.

Matsushima, T., Furuta, H., Matsuo, Y., Nachum, O., and Gu, S. Deployment-efficient reinforcement learning via model-based offline optimization. *ArXiv*, abs/2006.03647, 2020.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. *Proceedings of Machine Learning Research*, 48:1928–1937, 20-22 Jun 2016.

MuJoCo. Halfcheetah-v2. https://gym.openai.com/envs/HalfCheetah-v2/.

Mukherjee, S. and Maillard, O.-A. Distribution-dependent and time-uniform bounds for piecewise i.i.d bandits. *arXiv preprint arXiv:1905.13159*, 2019.

Murphy, S. A., van der Laan, M. J., and Robins, J. M. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 2001.

Nachum, O., Ahn, M., Ponte, H., Gu, S. S., and Kumar, V. Multi-agent manipulation via locomotion using hierarchical sim2real. *PMLR*, 100:110–121, 30 Oct–01 Nov 2020. URL http://proceedings.mlr.press/v100/nachum20a.html.

NCSS. Cumulative sum (cusum) charts. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/CUSUM_Charts.pdf.

Neyman, J., Pearson, E. S., and Pearson, K. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, 1933. doi: 10.1098/rsta.1933.0009.

O'Brien, P. C. and Fleming, T. R. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556, 1979. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2530245.

OpenAI. Pendulum-v0. https://gym.openai.com/envs/Pendulum-v0/.

Page, E. S. Continuous Inspection Schemes. *Biometrika*, 41(1-2):100–115, 06 1954. ISSN 0006-3444. doi: 10.1093/biomet/41.1-2.100. URL https://doi.org/10.1093/biomet/41.1-2.100.

Pardo, F., Tavakoli, A., Levdik, V., and Kormushev, P. Time limits in reinforcement learning. *CoRR*, abs/1712.00378, 2017. URL http://arxiv.org/abs/1712.00378.

PennState College of Science. Lecture notes in stat 509: Alpha spending function approach. https://online.stat.psu.edu/stat509/node/81/.

Petrov, V. V. *Sums of Independent Random Variables*. Nauka, 1972.

Pocock, S. J. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 08 1977. ISSN 0006-3444. doi: 10.1093/biomet/64.2.191. URL https://doi.org/10.1093/biomet/64.2.191.

Rockafellar, R. T. and Uryasev, S. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–41, 2000. doi: 10.21314/JOR.2000.038.

Ryan, T. P. *Statistical Methods for Quality Improvement*. Wiley; 3rd Edition, 2011.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012.

Wald, A. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186, 06 1945. doi: 10.1214/aoms/1177731118. URL https://doi.org/10.1214/aoms/1177731118.

Westgard, J., Groth, T., Aronsson, T., and Verdier, C. Combined shewhart-cusum control chart for improved quality control in clinical chemistry. *Clinical chemistry*, 23: 1881–7, 11 1977. doi: 10.1093/clinchem/23.10.1881.

Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 03 1938. doi: 10.1214/aoms/1177732360. URL https://doi.org/10.1214/aoms/1177732360.

Williams, S. M. et al. Quality control: an application of the cusum. *BMJ: British medical journal*, 304.6838:1359, 1992.

Xie, Y. and Siegmund, D. Weak change-point detection using temporal correlation, 2011.

Yashchin, E. On the analysis and design of cusum-shewhart control schemes. *IBM Journal of Research and Development*, 29(4):377–391, 1985.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization, 2020.

Zhao, X. et al. Assessing the safety and reliability of autonomous vehicles from road testing. *ISSRE*, 2019.

Zhou, S., Jin, N., and Jin, J. J. Cycle-based signal monitoring using a directionally variant multivariate control chart system. *IIE Transactions*, 37(11):971–982, 2005. doi: 10.1080/07408170590925553. URL `https://doi.org/10.1080/07408170590925553`.

Contents

List of Theorems

A. Detailed Preliminary Materials

A.1. Reinforcement Learning and Episodic Framework

The environment of a Reinforcement Learning (RL) problem is usually modeled as a *Decision Process*. This is essentially a state-machine, where the (possibly random) transition between states depends on decision-making, as well as on the current and the previous states (in the general case). Every state (and possibly every decision) is assigned a corresponding reward, and the goal of the decision-making system (termed *agent*) is to maximize some function of the rewards, named the *return function*. In contrast to Supervised Learning, the feedback from the environment does not inform the agent whether it succeeded to maximize the rewards, but merely how high the rewards were. It is up to the agent to explore the possible decisions (also termed *actions*) and the corresponding rewards.

The return function is usually a simple sum of the rewards for a finite process, and a decayed sum for an infinite process. In the finite case, the process usually repeats multiple times in different variants, e.g., with different initial states. Common examples are board and video games (Brockman et al., 2016), as well as more realistic problems such as repeating drives in autonomous driving task. In the context of RL, the repetitions of the decision process are usually named *episodes*. Bylander (Bylander) defined an episode as the ”path from initial to a terminal state”. Pardo et al. (Pardo et al., 2017) wrote that ”it is common to let an agent interact for a fixed amount of time with its environment before resetting it and repeating the process in a series of episodes”.

Note that once the agent chooses a decision-making scheme (termed *policy*), the decision process essentially reduces to a (decision-free) random process. Every time-step in the process has a certain distribution of (state and) reward, and different time-steps may depend on each other.

The decision process in RL is often modeled as a *Markov Decision Process* (MDP) (Bellman, 1957), where every state depends only on the preceding state and the agent's action. The decision-free process received from an MDP with relation to a fixed policy is a *Markov Chain* (MC), which under certain further assumptions is guaranteed to converge into a stationary state (Freedman, 2017). However, even in such a restrictive model, long-term correlations between rewards may still carry information if the states are not observable by the agent; and even under the further conditions of convergence to a stationary state, the rate of convergence may be slow compared to the length of an episode. The non-stationarity of the rewards within an episode is demonstrated, for example, in Fig. 1b.

This work exploits the repetitive nature of the episodic random processes – and in particular the rewards of episodic decision processes in the context of RL – to estimate the expectations and the correlations in the process. Since we measure the rewards directly, without considering the underlying states or any other observations available to the agent, we may call this approach model-free in the context of RL.

Note that in the scope of this work, the goal of the episodes is to provide i.i.d samples of a non-i.i.d random process, so that the covariance parameters of the process can be estimated. Hence, the scope of ”episodic problems” may be quite extensive: it may include even life-time systems that run continuously without ever resetting – as long as a reference dataset of other instances of the system is available, and the sample resolution does not introduce too many covariance parameters to estimate from the reference dataset. Indeed, the model defined in Section C and the optimality results in Section D are fully capable of handling a part of a single, long episode (with the exception of the asymptotic results in Section D.1.1).

A.2. Hypothesis Testing

In a standard hypothesis test, two hypotheses are formulated regarding some observable phenomenon, and we wish to decide which one is true according to available evidence, given in the form of observations $X \in \mathbb{X}$ from a corresponding observation space $\mathbb{X}$. One hypothesis is often regarded as the default, named the *Null Hypothesis* and denoted H_0; and given X we have to decide whether to *reject* H_0 in favor of the *Alternative Hypothesis* H_A.

The fundamental distinction between the hypotheses lays on their different probabilistic models $P\left(X \middle| H\right)$ (either probability function or probability density function), also referred to as the *likelihood* $L\left(H \middle| X\right)$ of the hypothesis given the observations. The difference between the models is often formulated in terms of different values of a parameter θ for some parametric probability function $P\left(X \middle| \theta\right)$. A *complex* hypothesis is one that allows different possible probabilistic models, represented by a set Θ of permitted values of θ. The likelihood of a complex hypothesis $H : \theta \in \Theta$ is defined as $L\left(H \middle| X\right) = \sup_{\theta \in \Theta} P\left(X \middle| \theta\right)$. The *likelihood-ratio* between two hypotheses is defined as $LR\left(H_0, H_A \middle| X\right) = \frac{L\left(H_0 \middle| X\right)}{L\left(H_A \middle| X\right)}$.

The log-likelihood-ratio is often used instead (Wilks, 1938), since it tends to derive simpler expressions for exponential families of distributions such as the Normal distribution. In this work we often denote $\lambda_{LR}\left(H_0, H_A \middle| X\right) = 2\ln\left(LR\right)$.

The basic metrics for the efficiency of a hypothesis test are its *significance* $P\left(\text{not reject } H_0 \middle| H_0\right) = 1 - P\left(\text{type-I error}\right) = 1 - \alpha$ and its *power* $P\left(\text{reject } H_0 \middle| H_A\right) = 1 - P\left(\text{type-II error}\right) = \beta$. A statistical hypothesis test with significance $1-\alpha$ and power β is said to be *optimal* if any statistical test with as high significance $1 - \tilde{\alpha} \geq 1 - \alpha$ has smaller power $\tilde{\beta} \leq \beta$.

According to Neyman-Pearson lemma (Neyman et al., 1933), a threshold-test on the likelihood ratio is an optimal hypothesis test. In a likelihood-ratio threshold-test with a threshold $\kappa \in \mathbb{R}$, we reject H_0 if $LR\left(H_0, H_A \middle| X\right) < \kappa$; reject with a certain probability $\rho \in [0,1]$ if $LR = \kappa$; and do not reject H_0 otherwise. Note that the behavior in the edge-case $LR = \kappa$ (controlled by ρ) only matters in the case of non-continuous distributions, where it is possible that $P\left(LR = \kappa\right) \neq 0$.

Note that the optimal hypothesis test is not unique, but rather leaves a degree of freedom in the tradeoff between α and β. In the case of a threshold-test, this degree of freedom is controlled by the threshold κ (and the edge probability ρ). It is common to define the test according to a desired significance level (often $\alpha = 0.01$ or $\alpha = 0.05$), and derive the corresponding threshold κ_α.

In certain cases, given a test-statistic and desired α, the threshold κ_α can be analytically calculated from the corresponding probabilistic model $P\left(X \middle| H_0\right)$. If the model is too complex or not well-defined, but expresses the sum of i.i.d random variables, then according to the *Central Limit Theorem* (CLT) (Petrov, 1972; Irwin, 2006), the model becomes closer to a Normal distribution as the number of summed variables grows, allowing to analytically calculate the asymptotic value of κ_α. Note that the CLT lays on the independence and identical distributions of the summed variables – two properties which are not generally satisfied by episodic rewards in the decision processes described in Section A.1.

Numeric methods are also available for estimation of properties of a hypothesis test (or the properties of a statistic of the observations). In *Monte-Carlo method* (Kroese et al., 2014), the test is simulated (or the statistic is computed) multiple times for observations X generated in a way which is assumed to be similar to a hypothesis H (in particular H_0 for significance estimation). In the *bootstrap* method (Efron, 2003), given i.i.d observations $X \in \mathbb{R}^n$ (assumed to be drawn according to a hypothesis H), Monte-Carlo method is applied on artificial observations X_b drawn by repeatedly sampling n elements from X with replacement.

A.3. Sequential Tests

Section A.2 describes the general scheme of a standard hypothesis test for distinction between two hypotheses according to certain available data. In many practical applications, the hypothesis test is repeatedly applied as the data change or grow, a procedure known as a *sequential test*. If the null hypothesis H_0 is true, and any individual hypothesis test falsely rejects H_0 with some probability α, then the probability that at least one of the multiple tests will reject H_0 is $\alpha_0 > \alpha$, termed *family-wise type-I error rate*. For simplicity, consider the private case of k independent tests, where $\alpha_0 = 1-(1-\alpha)^k \xrightarrow{k\to\infty} 1$.

This problem, also known as inflation of significance or inflation of α in sequential tests, was addressed by many over the years. A simple solution is the *Bonferroni correction* (Goldman, 2008), setting significance level of $1 - \alpha/k$ in every individual test. This way, we have $P(\exists i : \text{test } i \text{ rejects}|H_0) \leq \sum_{i=1}^k \alpha/k = \alpha$. However, the inequality becomes equality only if the rejections of the various tests are disjoint events (not even independent); thus in practice we often have $\alpha_0 \ll \alpha$, which makes the Bonferroni correction extremely conservative. Appendix B describes other relevant works on sequential testing.

B. Related Work: Detailed Discussion

As explained in Section 2, sequential tests repeatedly apply individual hypothesis tests with certain significance level $1 - \alpha \in (0,1)$. The probability that at least one test would reject the null hypothesis H_0 increases with the number of the individual tests, leading to "inflation of α" and decreased family-wise significance level $1 - \alpha_0 < 1 - \alpha$. Section 5 discusses this problem in the context of tests on episodic signals. Here we discuss some of the existing methods for design of sequential tests.

Sequential Probability Ratio Test (SPRT): SPRT (Wald, 1945) considers a symmetric approach between two hypotheses H_1,H_2, and aims to decide between them as fast as possible, subject to the probability of a wrong decision being bounded by α. The decision rule is chosen such that the expected time until decision is minimized. The element that bounds the probability of wrong decision is the setup of the flow of the test. Every iteration, the decision rule decides between three possibilities: accept H_1, accept H_2, or continue. The possibility to stop on acceptance of the true hypothesis limits the inflation of α.

In contrast to this setup, in the change-point detection problem – where continuously looking for changes – we either reject H_0 or continue, but never stop to accept H_0. Dedicated sequential tests are designed for the problem of change-point detection.

Cumulative Sum test (CUSUM): The CUSUM test (Page, 1954; NCSS) is a well-studied (Brook et al., 1972; Yashchin, 1985) and very popular method in quality control and change detection (Williams et al., 1992; Westgard et al., 1977). While being useful in a wide scope of problems, the test requires the size of change to be defined in advance as a parameter (a requirement that exists in other methods as well (Lund et al., 2007)). In addition, CUSUM assumes to observe i.i.d samples. The statistic is defined incrementally in a non-linear way, making it more difficult to generalize to non-i.i.d models, although several generalizations do exist, e.g., for the case of first-order autoregressive signal AR(1) (Lu & Jr., 2001). However, for example, Fig. 1c demonstrates empiric rewards in HalfCheetah environment (MuJoCo), where the dependencies in the signal require a more expressive model.

Persistent drift and Dickey-Fuller test: Certain methods are available for detection of persistent drifts (also known as trends) in time-series. For example, Dickey-Fuller test (Dickey & Fuller, 1979) for unit-roots in autoregressive models essentially looks for linear drifts. However, in the scope of this work we do not assume a persistent drift, nor limit ourselves to autoregressive models.

α-spending functions: The α-spending functions (Lan, 1994; PennState College of Science) deal with the inflation of α in sequential tests by conceptually referring to α as a limited budget of significance, where every individual test spends some of the budget. Due to the dependence between the individual tests, the total budget spent is smaller than the sum of the individual spends $\alpha_0 < \sum_i \alpha_i$. Thus, careful calculations are required for tuning of the family-wise significance level α_0.

Pocock (1977), for example, showed how to calculate a constant individual significance level $\alpha_i \equiv \alpha$ given a desired family-wise significance α_0 and known number of k individual tests, assuming that the tests are applied to accumulated normal i.i.d data samples. For many applications, such a constant significance level tends to spend too much α-budget in the first individual tests, reducing too much power from the later tests – where most of the data are available. It is often preferred to keep high significance level for these final tests, and reject H_0 in earlier tests only in radical cases. Accordingly, the O'Brien-Fleming function (O'Brien & Fleming, 1979) determines the individual significance levels $\{\alpha_i\}_{i=1}^k$ under similar i.i.d and normality assumptions as Pocock, but lets α_i gradually increase over the sequential test. In Section 5 we consider the α-spending approach and generalize it through a bootstrap mechanism to handle any sequence of individual tests for the case of episodic data; that is, i.i.d episodes consisting of samples which are not assumed to be independent, normal, or identically-distributed.

Multivariate mean shift: In a way, our work can be seen as a test for change-point or mean-shift of i.i.d T-dimensional multivariate random variables – the episodes. This problem was addressed before, e.g., using Hotelling statistic (Hotelling, 1931), histograms comparison (Boracchi et al., 2018), and K-L distance (Kuncheva, 2013). However, our setup has two essential differences from the multivariate mean-shift problem: first, since we look for a signed (negative) change in a univariate signal, we form the test's alternative hypothesis H_A correspondingly. This results in the uniform and partial degradation hypotheses, which are essentially different from the alternative hypothesis of Hotelling test, for example. Indeed, Section 6 demonstrates the advantage over Hotelling in the framework of RL, that is, episodic univariate rewards signal.

Second, since the episodic signal is *temporal* univariate, the coordinates of the "multivariate variables" are not observed simultaneously. As a result, when observing in the middle of an episode, we have incomplete information about the last multivariate variable (and possibly the first one, depending on how the lookback-horizon is defined). Both BFAR and the test statistics in this work take care of this issue. This is required for correct inference at any mid-episode time, but is particularly important for fast detection of large changes – which should be detected in the middle of the first episode.

Numeric methods: Colas et al. (2019) address the problem of comparing different RL algorithms, referring to whole episode as a single data sample for the tests. Harel et al. (2014) apply permutations test to detect changes in i.i.d data, focusing on drifts that impair predictive models of the data. The bootstrap mechanism discussed in Section 5 can be seen as a permutations test on i.i.d episodes (instead of single samples). Abhishek & Mannor (2017) also bring together ideas from bootstrap and sequential tests to construct a nonparametric sequential hypothesis test. The test applies bootstrap on single samples within blocks of data, assuming the data samples are i.i.d. Certain machine-learning based approaches were also suggested for changepoint detection in time-series (Aminikhanghahi & Cook, 2016). Ditzler et al. (2015) wrote that "change detection is typically carried out by inspecting independent and identically distributed (i.i.d) features extracted from the incoming data stream, e.g., the sample mean, the sample variance, and/or the classification error".

Changing environment and safety in RL: In Multi-Armed Bandits (MAB) (Berry & Fristedt, 1985), where by default each bandit (action) yields i.i.d rewards, several works address the problem of regret minimization (namely, optimization of rewards during training) with abrupt changes (Garivier & Moulines, 2011; Mukherjee & Maillard, 2019), gradual changes (Besbes et al., 2014) and even adversarial changes (Lykouris et al., 2020; Alatur et al., 2020; Gupta et al., 2019; Jun et al., 2018).

Training in presence of non-stationary environment was also considered in other environments such as multi-agent environments (Hernandez-Leal et al., 2019) and in model-based RL with varying model (Lecarpentier & Rachelson, 2019; Banerjee et al., 2016). Several works addressed the problem of safety in exploration of RL algorithms during training (Garcia & Fernandez, 2015; Chow et al., 2018; Junges et al., 2016), often using model-based learning of the environment (Cheng et al., 2019) or specified constraints (Alshiekh, 2017).

Note that our work refers to changes beyond the scope of the training phase, at the stage where the agent is fixed and required not to train further, in particular not in an online manner. Robust algorithms may prevent rewards degradation in the first place, but when they do not – it is crucial to be alerted. To the best of our knowledge, we are the first to exploit correlations between rewards in post-training phase to test for changes in both model-based and model-free RL.

C. Extended Definitions and Model Discussions

Episodic signal model: Below is the formal definition of an episodic signal, as discussed in Section C.

Definition C.1 (Episodic index decomposition)**.** *Let $t, T \in \mathbb{N}$. We define $k(t,T) := \lfloor \frac{t-1}{T} \rfloor$, $\tilde{\tau}(t,T) := t \,(mod\, T)$, and $\tau(t,T) := \begin{cases} T & \text{if } \tilde{\tau}(t,T) = 0 \\ \tilde{\tau}(t,T) & \text{if } \tilde{\tau}(t,T) \neq 0 \end{cases}$. When no confusion is risked, we may simply write $k = k(t,T), \tau = \tau(t,T)$. Note that $\forall t, T \in \mathbb{N} : t = kT + \tau$.*

Definition C.2 (T-long episodic signal; an extended formulation of Definition 4.1)**.** *Let $n, T \in \mathbb{N}$. Denote $K = k(n,T)$, $\tau_0 = \tau(n,T)$ according to Definition C.1. A sequence of real-valued random variables $\{X_t\}_{t=1}^n$ is a T-long episodic signal, if its joint probability density distribution can be written as*

$$f_{\{X_t\}_{t=1}^n}(x_1, ..., x_n) = \left[\prod_{k=0}^{K-1} f_{\{X_t\}_{t=1}^T}(x_{kT+1}, ..., x_{kT+T}) \right] \cdot f_{\{X_t\}_{t=1}^{\tau_0}}(x_{KT+1}, ..., x_{KT+\tau_0}) \quad (3)$$

(where in the edge case $K = 0$ we define the empty product to be 1). We further denote $\boldsymbol{\mu_0} := E[(X_1, ..., X_T)^\top] \in \mathbb{R}^T, \Sigma_0 := Cov((X_1, ..., X_T)^\top, (X_1, ..., X_T)) \in \mathbb{R}^{T \times T}$.

Expectation and covariance of an episodic signal: The expectations and covariance matrix of a whole episodic signal can be directly derived from the parameters $\boldsymbol{\mu_0}, \Sigma_0$ corresponding to the expectations and covariance matrix of a single episode.

Proposition C.1 (Expectation and covariance of an episodic signal)**.** *Let $\{X_t\}_{t=1}^n$ be a T-long episodic signal with parameters $\boldsymbol{\mu_0}, \Sigma_0$. The expectations $\boldsymbol{\mu} := E[(X_1, ..., X_n)^\top] \in \mathbb{R}^n$ and covariance matrix $\Sigma := Cov((X_1, ..., X_n)^\top, (X_1, ..., X_n)) \in \mathbb{R}^{n \times n}$ are uniquely determined by $\boldsymbol{\mu_0}$ and Σ_0, respectively.*

Proof. For any $t \in \{1, ..., n\}$, denote $t = kT + \tau$ according to Definition C.1. From Eq. (1) it is clear that $\forall t_1 =$

$k_1T+\tau_1, t_2=k_2T+\tau_2 \in \{1,...,n\}$:

$$\mu_{t_1} = E[X_{t_1}] = (\boldsymbol{\mu_0})_{\tau_1}$$
$$\Sigma_{t_1t_2} = \text{Cov}(X_{t_1}, X_{t_2}) = \begin{cases} (\Sigma_0)_{\tau_1\tau_2} & \text{if } k_1 = k_2 \\ 0 & \text{if } k_1 \neq k_2 \end{cases} \tag{4}$$

$\square$

Proposition C.1 essentially means that $\boldsymbol{\mu}$ consists of periodic repetitions of $\boldsymbol{\mu_0}$, and Σ consists of copies of Σ_0 as $T \times T$ blocks along its diagonal. For both parameters, the last repetition is cropped if $\tau(n,T) < T$.

Multivariate normal episodic signal: Some of the theoretical results in Section D assume multivariate normality of the episodic signal. The formal definition of such a signal is given below.

Definition C.3 (Multivariate normal T-long episodic signal)**.** *Let $\{X_t\}_{t=1}^n$ be a T-long episodic signal (Definition C.2). For any $1 \leq \tau \leq min(T,n)$, define $\boldsymbol{\mu_\tau} \in \mathbb{R}^\tau$ to be the first τ elements of $\boldsymbol{\mu_0}$ and $\Sigma_\tau \in \mathbb{R}^{\tau\times\tau}$ to be the upper-left $\tau \times \tau$ block of Σ_0. The signal $\{X_t\}_{t=1}^n$ is multivariate normal if $\forall 1 \leq \tau \leq min(T,n)$,*

$$f_{X_1,...,X_\tau}(\boldsymbol{x}) = (2\pi)^{-\tau/2} det(\Sigma_\tau)^{-1/2} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_\tau})^\top \Sigma_\tau^{-1}(\boldsymbol{x}-\boldsymbol{\mu_\tau})} \tag{5}$$

From Definitions C.2,C.3 it is clear that if $\{X_t\}_{t=1}^n$ form a multivariate normal T-long episodic signal, then in particular $X = (X_1,...,X_n)^\top \in \mathbb{R}^n$ is an n-dimensional multivariate normal variable, with expectations $\boldsymbol{\mu}$ and covariance Σ determined by Eq. (4).

Parameters estimation: As mentioned above, a possible way to estimate the parameters $\boldsymbol{\mu_0}, \Sigma_0$ of an episodic signal is to compute the mean vector and the covariance matrix of a dataset $\{x_{i\tau} | 1 \leq i \leq N, 1 \leq \tau \leq T\}$ of N episodes assumed to satisfy Eq. (1). According to the Central Limit Theorem (Petrov, 1972; Irwin, 2006), since the episodes are i.i.d, for any time-step τ the estimate $(\hat{\boldsymbol{\mu_0}})_\tau = \frac{1}{N}\sum_{i=1}^N x_{i\tau}$ is asymptotically normally-distributed around the true mean $(\boldsymbol{\mu_0})_\tau$ with variance $\frac{\text{Var}((\boldsymbol{\mu_0})_\tau)}{N}$. Furthermore, in the private case of a multivariate normal signal, the covariance matrix estimate $(\hat{\Sigma_0})_{ij} = \frac{1}{N-1}\sum_{k=1}^N (x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)$ follows Wishart distribution (K. V. Mardia & Bibby, 1979) (up to a factor of $N-1$), with $N-1$ degrees of freedom and variance $\text{Var}((\hat{\Sigma_0})_{ij}) = \frac{1}{N-1}\left((\Sigma_0)_{ii}(\Sigma_0)_{jj} + (\Sigma_0)_{ij}^2\right)$.

If N is suspected to be too small for accurate estimation, it is possible to deal with the estimation error of the model parameters through regularization. One possible regularization is assuming absence of correlations between distant time-steps ($\exists \delta \in \mathbb{N}, \forall |t_2-t_1| > \delta : (\Sigma_0)_{t_1t_2} = 0$). Another is to essentially reduce T through grouping of sequences of time-steps together (as we do in Section 6, for example).

In the analysis in the following sections we assume that both $\boldsymbol{\mu_0}$ and Σ_0 are known.

Multidimensional signals: For simplicity of the theoretical discussion, we only consider one-dimensional signals: for any t, the random variable X_t returns a scalar $x_t \in \mathbb{R}$. However, a generalization to multidimensional signals ($x_t \in \mathbb{R}^d$) is straight-forward: A d-dimensional T-long episodic signal is simply a one-dimensional (dT)-long episodic signal, where the observations arrive in groups of d samples per group (i.e., n is always an integer multiplication of d). Since the various dimensions are equivalent to time-steps in the eyes of this model, the correlations between the various dimensions are inherently captured. Note that for a large number of dimensions, the $\mathcal{O}(d^2T^2)$ degrees of freedom in the model may be impractical to estimate through a reference dataset.

D. Likelihood-Ratio Test for Drift in Episodic Signal: Formal Development

In this section we look for an optimal hypothesis test for detection of a negative drift in multivariate normal episodic signal (see Definitions C.2,C.3). The corresponding hypotheses are episodic signal with known parameters (H_0), and episodic signal with identical covariance matrix but smaller expected values (H_A), as defined below. By "optimal test" we mean that given the test's significance level (i.e., type-I error rate), it should provide the maximum possible power (i.e., minimum type-II error rate) with respect to H_A. To that end, we calculate the log-likelihood-ratio and use it (up to a monotonous transformation) as a test-statistic according to Neyman-Pearson lemma (Neyman et al., 1933).

In Section D.1.1, after proving optimality for a certain negative drift, we eliminate the multivariate-normality assumption and analyze the asymptotic power of the suggested statistical test. In particular, we show that it is asymptotically superior to a simple threshold-test on the average reward.

Note that in the scope of this section we assume an individual test at a certain point of time. Adjustment of the significance level to sequential tests is handled in Section 5.

Formally, the test is defined with respect to some real-valued random variables $X_1, ..., X_n$.

Definition D.1 (Null hypothesis)**.** *Let $\{X_t\}_{t=1}^n$ be real-valued random variables, and let $T \in \mathbb{N}, \boldsymbol{\mu_0} \in R^T, \Sigma_0 \in R^{T\times T}$. The null hypothesis $H_0(T, \boldsymbol{\mu_0}, \Sigma_0)$ in the scope of this section, is that $\{X_t\}_{t=1}^n$ form a T-long episodic signal (Definition C.2), with known parameters $T, \boldsymbol{\mu_0}, \Sigma_0$. For simplicity, we further assume that Σ_0 is of full-rank (i.e., invertible).*

We define a standard setup for most of the analysis below, both with and without the multivariate-normality assumption.

Definition D.2 (The standard setup)**.** *In the standard setup, we denote by $X = \{X_t\}_{t=1}^n$ a T-long episodic signal for some $n, T \in \mathbb{N}$ (Definition C.2), and let the null hypothesis H_0 be as in Definition D.1, with known parameters $\boldsymbol{\mu_0} \in \mathbb{R}^T, \Sigma_0 \in \mathbb{R}^{T\times T}$.*

Note that under H_0, the complete signal's expectations $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$ are also known through Proposition C.1.

We also denote $k(t) = k(t, T), \tau(t) = \tau(t, T)$ as in Definition C.1, and in particular $K := k(n, T), \tau_0 := \tau(n, T)$.

Definition D.3 (The standard normal setup)**.** *The standard normal setup is the standard setup where X is a multivariate-normal episodic signal (Definition C.3).*

D.1. Uniform Degradation Test

The general alternative hypothesis we use assumes conservation of the correlations structure of H_0, along with decrease in the expectations.

Definition D.4 (General degradation hypothesis)**.** *Given the standard setup (Definition D.2), let $\mathbb{E} \subseteq \mathbb{R}^T$ s.t. $\forall \boldsymbol{\epsilon_0} \in \mathbb{E}, 1 \leq t \leq T : (\boldsymbol{\epsilon_0})_t \geq 0$. According to the $\mathbb{E}$-degradation hypothesis, denoted $H_A(\mathbb{E})$, there exists $\boldsymbol{\epsilon_0} \in \mathbb{E}$ such that $\{X_t\}_{t=1}^n$ form $\tilde{T}$-long episodic signal with the parameters $\tilde{T} = T, \tilde{\Sigma_0} = \Sigma_0$ and $\tilde{\boldsymbol{\mu_0}} = \boldsymbol{\mu_0} - \boldsymbol{\epsilon_0}$.*

In particular, according to Eq. (4), the covariance and the mean of the whole signal under $H_A(\mathbb{E})$ are $\tilde{\Sigma} = \Sigma$ and $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} - \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}(\boldsymbol{\epsilon_0}) \in \mathbb{R}^n$ is a cyclic completion defined by $\forall t = kT + \tau : (\boldsymbol{\epsilon})_t := (\boldsymbol{\epsilon_0})_\tau$.

Proposition E.1 calculates the log-likelihood-ratio with respect to the hypotheses in Definitions D.1,D.4, assuming a multivariate-normal episodic signal. Still, to derive a concrete statistical test, further assumptions must be applied on $\mathbb{E}$. We begin with the *uniform degradation* assumption, corresponding to a disturbance source that affects the whole signal uniformly. For example, in the context of Reinforcement Learning, such a model may refer to changes in constant costs or action costs, as well as certain environment dynamics whose change influences the various states in a similar way.

Definition D.5 (Uniform degradation hypothesis)**.** *Let $\epsilon_0 > 0$. The uniform degradation hypothesis, denoted $H_A^{unif}(\epsilon_0)$, is a degradation hypothesis $H_A(\mathbb{E})$ with $\mathbb{E} := \{\epsilon \cdot \mathbf{1} | \epsilon \geq \epsilon_0\}$, where $\mathbf{1} := (1, ..., 1)^\top \in \mathbb{R}^T$.*

Fig. 2 demonstrates the empiric degradation in the rewards of a trained agent in HalfCheetah environment, following changes in gravity, mass, and control-cost (see Table 2 for details). It seems that some modifications indeed cause a quite uniform degradation, while in others the degradation is mostly restricted to certain ranges of time. This may be important, in particular if the non-degraded time-steps happen to be assigned large weights by the test, as demonstrated in Section 6.2. In Section D.2 we suggest an alternative model, whose corresponding test is proved in Section 6.2 to be more robust to such non-uniform degradation.

We now show that an optimal hypothesis test for detection of uniform degradation in multivariate normal episodic signal is a threshold-test on the weighted-mean of the signal, where the weights are derived from the inverted covariance matrix.

Note that according to Proposition C.1, the covariance matrix $\Sigma = \Sigma(\Sigma_0) \in \mathbb{R}^{n\times n}$ of the full signal is block-diagonal with the blocks being $\Sigma_0 \in \mathbb{R}^{T\times T}$ (or an upper-left block of Σ_0). Hence, the inverted Σ is given directly by inverting Σ_0 (and possibly one upper-left block of Σ_0).

Definition D.6 (Uniform degradation weighted-mean)**.** *Given the standard setup (Definition D.2), the uniform-degradation weighted-mean of X is $s_{unif}(X|\Sigma_0) := W \cdot X$, where $W := \mathbf{1}^\top \cdot \Sigma^{-1} \in \mathbb{R}^n$.*

Note that the first KT elements of W are T-periodic with $\forall 0 \leq k \leq K-1 : w_{kT+1}, ..., w_{kT+T} = \mathbf{1}^\top \cdot \Sigma_0^{-1} \in \mathbb{R}^T$. We define accordingly $W_0 := \mathbf{1}^\top \cdot \Sigma_0^{-1}$ and $W_{\tau_0} := (w_{KT+1}, ..., w_{KT+\tau_0})^\top = \mathbf{1}^\top \cdot \Sigma_{\tau_0}^{-1}$, where Σ_{τ_0} is the upper-left $\tau_0 \times \tau_0$ block of Σ_0.

Proposition E.3 shows the consistency of the uniform-degradation weighted-mean, and Theorem D.1 shows that it derives an optimal hypothesis test for uniform degradation.

Definition D.7 (Threshold test)**.** *Assume the standard setup (Definition D.2), and let $S : \mathbb{R}^n \to \mathbb{R}$ (statistic), $\kappa \in \mathbb{R}$ (threshold) and $\rho \in [0, 1]$ (edge-case probability). The corresponding κ-threshold-test is defined as follows:*

Given the observations $\forall 1 \leq t \leq n : X_t = x_t \in \mathbb{R}$, calculate the statistic $s = S(x_1, ..., x_n)$. If $s < \kappa$, reject H_0. If $s = \kappa$, reject H_0 with probability $p = \rho$ (note that this is only relevant for non-continuous distributions, where $P(S = \kappa) \neq 0$). If $s > \kappa$, do not reject H_0.

We denote the significance level of the test $\alpha := P(\text{reject } H_0 | H_0)$. For simplicity, in the discussion below we often omit ρ, implicitly assuming continuous distribution of the signal.

Theorem D.1 (Optimal test for uniform degradation; an extended formulation of Theorem 4.1)**.** *Assume the standard normal setup (Definition D.3) with $H_A^{unif}(\epsilon_0)$ of Definition D.5 as the alternative hypothesis, and let $\alpha \in (0, 1)$. Then, there exists $\kappa \in \mathbb{R}$ such that a κ-threshold-test on the uniform-degradation weighted-mean statistic has the greatest power among all the statistical tests with significance level $\tilde{\alpha} \leq \alpha$.*

Proof. The proof is available in Appendix E. Roughly speaking, according to Neyman-Pearson lemma (Neyman et al., 1933) a threshold-test on the likelihood-ratio is optimal, hence it is sufficient to show that the uniform-degradation weighted-mean s_{unif} is monotonous with the likelihood-ratio.

Note that the likelihood-ratio is taken with respect to a complex hypothesis $H_A^{unif}(\epsilon_0)$ that has a degree of freedom $\epsilon \in [\epsilon_0, \infty)$, where ϵ depends on X. Some algebraic work is required to show that ϵ only depends on X through s_{unif}, and that the whole likelihood-ratio is monotonous with s_{unif}. $\square$

Algorithm 3 describes the threshold-test in the non-sequential framework. The uniform-degradation test-statistic (or any other function) can be fed into the algorithm as an input.

As can be seen, the rejection threshold $\kappa_\alpha \in \mathbb{R}$ is chosen according to the desired type-I error rate $\alpha \in (0, 1)$, using a bootstrap mechanism described in Algorithm 2. B bootstrap-samples are sampled from a reference dataset of N episodes of the signal, assumed to follow the null hypothesis H_0 of Definition D.1. For each bootstrap-sample[1] the test-statistic is calculated, yielding a bootstrap-estimate for the distribution of the statistic under H_0. The rejection threshold κ_α is set to be the α-quantile of the estimated distribution. If the estimated distribution is close to the true distribution, then we have $P\left(s \leq \kappa_\alpha | H_0\right) \approx P\left(s \leq q_\alpha(s|H_0) | H_0\right) = \alpha$, where $q_\alpha(s|H_0)$ is the α-quantile of s under H_0.

D.1.1. Asymptotic analysis in absence of the normality assumption

The optimality of the uniform-degradation weighted-mean test (proved in Theorem D.1) relies on the assumption that the episodic signal is multivariate normal. In this section we show that even in absence of the normality assumption, the test while not necessary is asymptotically superior to a standard threshold-test on the average of the signal (though it is not necessarily the optimal test anymore).

Since the episodes in the signal are still assumed to be i.i.d, both a simple mean and the uniform-degradation weighted-mean s_{unif} are asymptotically normal (where $n \to \infty$ with respect to a constant episode length $T \in \mathbb{N}$). For simplicity of the asymptotic analysis below, we focus on integer number of episodes, i.e., $n = KT$ and $K \to \infty$ (rather than $n \to \infty$).

[1] As a terminological note, this sampling mechanism can be considered a bootstrap in the sense of distribution estimation from a single dataset using sampling with replacement; or can be merely considered a Monte-Carlo simulation in the sense that the test signal is compared to distribution estimated by an external simulative source (the reference data).

We also define normalized variants of our statistics, with zero-mean and unit-variance per episode:

$$
\begin{aligned}
s_{simp}(\{X_t\}_{t=1}^{n}) &:= \sum_{t=1}^{n} X_t \\
\tilde{s}_{simp}^{K} &:= \frac{s_{simp}(\{X_t\}_{t=1}^{KT}) - K \cdot E\left[s_{simp}(\{X_t\}_{t=1}^{T}) \middle| H_0\right]}{\sqrt{K \cdot \text{Var}(s_{simp}(\{X_t\}_{t=1}^{T}) \middle| H_0)}} \\
\tilde{s}_{unif}^{K} &:= \frac{s_{unif}(\{X_t\}_{t=1}^{KT}) - K \cdot E\left[s_{unif}(\{X_t\}_{t=1}^{T}) \middle| H_0\right]}{\sqrt{K \cdot \text{Var}(s_{unif}(\{X_t\}_{t=1}^{T}) \middle| H_0)}}
\end{aligned}
\tag{6}
$$

Note that Algorithm 3 is invariant to linear transformation of the statistic, since the test-statistic and the reference bootstrap distribution pass through the same transformation. Hence, the tests on s_{simp}, s_{unif} are equivalent to the tests on $\tilde{s}_{simp}, \tilde{s}_{unif}$, respectively.

Since $\tilde{s}_{simp}, \tilde{s}_{unif}$ are asymptotically normal with zero-mean and unit-variance under H_0, the desired test threshold for sufficiently large K is $\kappa \approx q_\alpha^0$, where q_α^0 is the α-quantile of the standard normal distribution. This threshold should be indirectly estimated by Algorithm 2.

Note that the sequential test of Algorithm 5 in Section 5 applies the individual tests of Algorithm 3 on a constant number of episodes (defined by the lookback horizon h). Hence, in the context of the sequential tests suggested in this work, the asymptotic analysis in this section refers to a very long lookback horizon, rather than very long running time. Regardless, as the analysis refers to a varying n, we need to generalize the standard setup (that assumes a constant signal length n).

Definition D.8 (The rolling setup)**.** *Let $\{X_t\}_{t\in\mathbb{N}}$ be an infinite sequence of real-valued random variables. In the rolling setup, for any $n \in \mathbb{N}$ we assume the standard setup of Definition D.2 with relation to the variables $\{X_t\}_{t=1}^{n}$ and the parameters $T, \boldsymbol{\mu_0}, \Sigma_0$ (which are independent of n).*

We first show that the test threshold $\kappa = q_\alpha^0$ indeed yields asymptotic significance level of $1-\alpha$, and guarantees asymptotic rejection of H_0 for uniform degradation of any size $\epsilon > 0$. Note that Algorithm 3 does not pick q_α^0 directly as a threshold, but should estimate it indirectly through Algorithm 2.

Proposition D.1 (Uniform degradation test consistency)**.** *Given the rolling setup, we define the alternative hypothesis H_A^ϵ to be that $\forall K \in \mathbb{N}$, the parameters of the signal $\{X_t\}_{t=1}^{KT}$ are $\boldsymbol{\mu_0} - \epsilon \cdot \mathbf{1}, \Sigma_0$ (i.e., $H_A(\{\epsilon\mathbf{1}\})$ in terms of Definition D.4). Given a significance level $\alpha \in (0,1)$, we have*

$$
\begin{aligned}
lim_{K\to\infty} P\left(s \le q_\alpha^0 \middle| H_0\right) &= \alpha \\
lim_{K\to\infty} P\left(s \le q_\alpha^0 \middle| H_A^\epsilon\right) &= 1
\end{aligned}
$$

for both $s = \tilde{s}_{simp}^K$ and $s = \tilde{s}_{unif}^K$ of Eq. (6), where q_α^0 is the α-quantile of the standard normal distribution.

Proof. The proof, fully provided in Appendix E, applies the Central Limit Theorem (Petrov, 1972; Irwin, 2006) on the i.i.d episodes. □

Theorem D.2 quantifies the asymptotic power of the threshold test for both simple mean and uniform-degradation weighted-mean. To that end, we consider uniform-degradation scaled as $\epsilon \propto \frac{1}{\sqrt{K}}$. We also denote by Φ the Cumulative Distribution Function of the standard normal distribution

Theorem D.2 (Uniform degradation test asymptotic power; an extended formulation of Theorem 4.2)**.** *Given the rolling setup, we define the alternative hypothesis $H_A^{\epsilon,K}$ to be that $\forall K \in \mathbb{N}$, the parameters of the signal $\{X_t\}_{t=1}^{KT}$ are $\boldsymbol{\mu_0} - \frac{\epsilon}{\sqrt{K}} \cdot \mathbf{1}, \Sigma_0$ (i.e., $H_A(\{\frac{\epsilon}{\sqrt{K}}\mathbf{1}\})$ of Definition D.4). Given a significance level $\alpha \in (0,1)$, we have*

$$
\begin{aligned}
lim_{K\to\infty} P\left(\tilde{s}_{simp}^K \le q_\alpha^0 \middle| H_A^{\epsilon,K}\right) &= \Phi\left(q_\alpha^0 + \frac{\epsilon T}{\sqrt{\mathbf{1}^\top \Sigma_0 \mathbf{1}}}\right) \\
\le \Phi\left(q_\alpha^0 + \epsilon\sqrt{\mathbf{1}^\top \Sigma_0^{-1} \mathbf{1}}\right) &= lim_{K\to\infty} P\left(\tilde{s}_{unif}^K \le q_\alpha^0 \middle| H_A^{\epsilon,K}\right)
\end{aligned}
$$

Proof. Similarly to Proposition D.1, the proof applies the Central Limit Theorem on the i.i.d episodes to calculate the asymptotic properties. Full details are provided in Appendix E. □

Note that while Theorem D.1 shows optimality of the uniform-degradation weighted-mean test for multivariate-normal episodic signal, Theorem D.2 proves that even in absence of normality the test is asymptotically superior to a threshold-test on the simple mean.

Finally, we quantify the difference of power between the tests.

Definition D.9 (Uniform degradation asymptotic power gain)**.** *Given the rolling setup, the uniform-degradation power gain is defined to be* $G^2 := \frac{(\mathbf{1}^\top \Sigma_0^{-1}\mathbf{1})(\mathbf{1}^\top \Sigma_0 \mathbf{1})}{T^2}$.

Note that according to Theorem D.2, if the asymptotic power of the simple-mean threshold-test with relation to the alternative hypothesis $H_A^{\epsilon,K}$ *is* $\Phi(q_\alpha^0 + y)$ *(where* $y \in \mathbb{R}$*), then the asymptotic power of the weighted-mean threshold-test is* $\Phi(q_\alpha^0 + G \cdot y)$.

Proposition D.2 (Uniform degradation test asymptotic power gain)**.** *Under the setup of Theorem D.2, there exist positive weights* $\{w_{ij}\}_{i,j=1}^T$ *such that the uniform-degradation power gain is*

$$G^2 = 1 + \sum_{i,j=1}^{T} w_{ij}(\lambda_i - \lambda_j)^2$$

where $\{\lambda_i\}_{i=1}^T$ *are the eigenvalues of* Σ_0.

Proof. The result is received from simple algebra after diagonalizing the symmetric positive-definite covariance matrix Σ_0. The full details are available in Appendix E. □

Clearly, the asymptotic power gain G becomes larger as the covariance matrix Σ_0 introduces more heterogeneous eigenvalues. Note that in the independent case, the eigenvalues are simply the variances of the different time-steps. In particular, in the i.i.d case, the variances are identical and the gain becomes $G = 1$, which is consistent with the fact that the two tests are equivalent in this case.

D.2. Partial Degradation Test

Definition D.5 assumes uniform degradation over all the time-steps in every episode. However, the effects of many environmental changes may focus on certain states (which is translated in our model-free setup into "certain time-steps"). An example is available in Fig. 2, as discussed before. To model such effects we introduce the *partial degradation hypothesis*.

Definition D.10 (Partial degradation hypothesis)**.** *Let* $\epsilon > 0, p \in (0,1)$*. Define* $A_m^T := \{a_0 \in \{0,1\}^T | \sum_{t=1}^T (a_0)_t = m\}$ *the set of binary vectors with exactly* m *one-entries. The partial degradation hypothesis, denoted* $H_A^{part}(\epsilon, p)$*, is a degradation hypothesis* $H_A(\mathbb{E})$ *(see Definition D.4) with* $\mathbb{E} := \{\epsilon \cdot a_0 | a_0 \in A_{\lceil pT \rceil}^T\}$.

Interpretation: As a private case of Definition D.4, Definition D.10 assumes conservation of the correlations structure. One possible interpretation of this assumption is causal relationships (where change in a certain time-step affects any other time-steps correlated with it). Another possible interpretation is that the partial degradation hypothesis does not restrict the degradation to only $m = \lceil pT \rceil$ time-steps, but rather distributes the degradation from m time-steps to all the episode, according to the same relations that created the correlations in the signal from the first place.

Similarly to the case of uniform-degradation, we can use the likelihood-ratio to derive a test-statistic and prove its approximate optimality with respect to $H_A^{part}(\epsilon, p)$.

Definition D.11 (Partial degradation mean)**.** *Given the standard setup (Definition D.2), denote* $\tilde{X} := X - \mu$*,* $\tilde{s} := \Sigma^{-1}\tilde{X} \in \mathbb{R}^n$ *and* $\forall : 1 \le \tau \le T : \tilde{S}_\tau := \sum_{k=0}^{\lfloor \frac{n-\tau}{T} \rfloor} \tilde{s}_{kT+\tau}$*. Given* $a_0 \in \{0,1\}^T$ *denote* $o(a_0) := \{1 \le t \le T | (a_0)_t = 1\}$*. Given* $p \in (0,1)$*, the* p*-partial-degradation mean of* X *is* $s_{part}(X;p) := min_{a_0 \in A_m^T} \sum_{\tau \in o(a_0)} \tilde{S}_\tau$*, where* $m(p) = \lceil pT \rceil$ *and* A_m^T *is defined as in Definition D.10.*

Note that while a has $|A_m^T| = \binom{T}{m}$ possible values, to compute s_{part} we only need to sum the lowest m values in $\{\tilde{S}_\tau\}_{\tau=1}^T$.

Theorem D.3 (Optimal test for partial degradation)**.** *Consider the standard normal setup (Definition D.3) with $H_A^{part}(\epsilon, p)$ of Definition D.10 as the alternative hypothesis, and let $\alpha \in (0, 1)$. Denote by P_α the largest possible power (with respect to $H_A^{part}(\epsilon, p)$) of a statistical test with significance level $\leq \alpha$. Then, there exists $\kappa \in \mathbb{R}$ such that the power of the κ-threshold-test on the partial-degradation mean is $P_\alpha - \mathcal{O}(\epsilon)$ (where $\mathcal{O}(\epsilon)$ is defined with relation to $\epsilon \to 0$).*

Proof. The proof is provided in Appendix E. Similarly to the proof of Theorem D.1, it is based on calculation of the log-likelihood-ratio λ_{LR} from Lemma 1 – after substituting Definition D.10. The calculation results in s_{part} along with an ϵ-dependent term whose effect on the test power is shown to be $\mathcal{O}(\epsilon)$. □

The parameter p and comparison to CVaR: In Definition D.11, the "weighted mean" completely eliminates $T - m$ of the entries of $\Sigma^{-1}X$, and only sums the most negative ones. Hence it can be interpreted as the well known CVaR statistic (Rockafellar & Uryasev, 2000) of $X - \mu$ after transformation to Σ^{-1}-basis. CVaR (Conditional Value at Risk) is intended to measure the "risky" tail of a random variable's distribution (Chen, 2020) by estimating its expectation – conditioned on it being below the p-quantile of the distribution. This is done simply by averaging the p "worst" (lowest) values in the corresponding data. To express risk, the parameter p is often set below 5% (MathWorks).

In our context, however, the relative part of time-steps p represents the scope of degradation rather than extremity of risk, and there is usually no reason to assume that $p \ll 1$. Such an assumption, when misplaced, may cause elimination of necessary information from the statistic. In fact, $0.9 \leq p < 1$ is shown in Section 6 to often achieve superior results, presumably because it maintains most of the information while still being able to filter noisy or misleading time-steps (e.g., time-steps with particularly large values).

Note that filtering positive time-steps is not "cheating" in the context of our problem: we essentially apply a monitor which looks for negative changes, and thus positive changes in other time-steps are indeed considered as noise for the sake of our monitor. If a more symmetric comparison is desired, then the test can be applied twice – once for negative changes, and once for positive changes.

The dependence on ϵ and $\mathcal{O}(\epsilon)$ approximation: The partial degradation mean is shown to be equal to an optimal test-statistic up to $\mathcal{O}(\epsilon)$. The approximation is used to handle the dependence of the minimum

$$\min_{a_0 \in A_m^T} \left(a^\top \Sigma^{-1} \tilde{X} + 0.5\epsilon (a^\top \Sigma^{-1} a) \right)$$

on ϵ. Note that for small degradation the second term is indeed negligible, while for larger degradation the distinction between the two hypotheses should pose little challenge to any detection algorithm.

Furthermore, the test is entirely invariant to a constant additive factor (due to the adjustment of the test threshold using the bootstrap in Algorithm 2); hence, the true distortion in the test is not of size $\gamma = 0.5\epsilon(a^\top \Sigma^{-1} a)$, but rather the change in γ due to the possibly-changed choice of a. Note that (a) since Σ^{-1} is positive-definite, we have $\forall a : a^\top \Sigma^{-1} a > 0$, hence the change in γ is necessarily smaller than γ; (b) if the parameter p is close to 100% (as discussed above), then most of the entries of a are necessarily kept unchanged, further reducing the change in γ.

If we wish to apply a more formal test, we can define for example $H_A^{part}(p) := \exists \epsilon > 0 : H_A^{part}(\epsilon, p)$ (similarly to Definition D.5 of uniform degradation, for $\epsilon_0 = 0$), which yields the log-likelihood-ratio $\min_{a \in A_m^T} -\frac{(a^\top \Sigma^{-1} \tilde{X})^2}{a^\top \Sigma^{-1} a}$ s.t. $a^\top \Sigma^{-1} \tilde{X} \leq 0$ (after minimization with relation to $\epsilon > 0$, similarly to the proof of Theorem D.1). Due to the discrete domain A_m^T of a, this becomes a non-linear integer programming problem, which should be solved for every run of the test on new data X.

Note that from a practical point of view, a major role of the partial-degradation model is to allow focusing on negative (degrading) entries of $a^\top \Sigma^{-1} \tilde{X}$, and filtering positive ones (as discussed above). For this role, both the approximate s_{part} and the accurate minimizer of $H_A^{part}(p) = \exists \epsilon > 0 : H_A^{part}(\epsilon, p)$ above are perfectly qualified, as both would tend to reject positive entries of $a^\top \Sigma^{-1} \tilde{X}$.

In the scope of this work, we stick to the approximately-correct and computationally-simpler partial degradation mean s_{part}.

E. Supplementary Calculations

Proposition E.1 (Likelihood ratio with respect to general degradation)**.** *Let the standard normal setup (Definition D.3 with the $\mathbb{E}$-degradation hypothesis $H_A(\mathbb{E})$ (Definition D.4). Define $\boldsymbol{\epsilon}(\boldsymbol{\epsilon_0})$ as in Definition D.4, and denote $\tilde{X}_t := X_t - (\boldsymbol{\mu})_t$. Then, the log-likelihood $\lambda_{LR}(H_0, H_A|\{X_t\}_{t=1}^n) := 2ln(\frac{P(\{X_t\}_{t=1}^n|H_0)}{sup_{H\in H_A} P(\{X_t\}_{t=1}^n|H)})$ of $\{X_t\}_{t=1}^n$ with respect to (the simple hypothesis) H_0 and (the complex hypothesis) H_A is*

$$\begin{aligned} &\lambda_{LR}(H_0, H_A|\{X_t\}_{t=1}^n) = \\ &\quad min_{\boldsymbol{\epsilon_0}\in\mathbb{E}} 2(\boldsymbol{\epsilon}(\boldsymbol{\epsilon_0}))^\top \Sigma^{-1}\tilde{X} + (\boldsymbol{\epsilon}(\boldsymbol{\epsilon_0}))^\top \Sigma^{-1}\boldsymbol{\epsilon}(\boldsymbol{\epsilon_0}) \end{aligned} \tag{7}$$

Proof. Using the density function of Eq. (5), we have

$$\begin{aligned} &\lambda_{LR}(H_0, H_A|\{X_t\}_{t=1}^n) = \\ &= 2\ln\left(\frac{e^{-0.5\tilde{X}^\top\Sigma^{-1}\tilde{X}}}{\max_{\boldsymbol{\epsilon_0}\in\mathbb{E}} e^{-0.5(\tilde{X}+\boldsymbol{\epsilon})^\top\Sigma^{-1}(\tilde{X}+\boldsymbol{\epsilon})}}\right) = \\ &= \min_{\boldsymbol{\epsilon_0}\in\mathbb{E}} (\tilde{X}+\boldsymbol{\epsilon})^\top\Sigma^{-1}(\tilde{X}+\boldsymbol{\epsilon}) - \tilde{X}^\top\Sigma^{-1}\tilde{X} \\ &= \min_{\boldsymbol{\epsilon_0}\in\mathbb{E}} \boldsymbol{\epsilon}^\top\Sigma^{-1}\tilde{X} + \tilde{X}^\top\Sigma^{-1}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\top\Sigma^{-1}\boldsymbol{\epsilon} \\ &= \min_{\boldsymbol{\epsilon_0}\in\mathbb{E}} 2\boldsymbol{\epsilon}^\top\Sigma^{-1}\tilde{X} + \boldsymbol{\epsilon}^\top\Sigma^{-1}\boldsymbol{\epsilon} \end{aligned}$$

where the last equality relies on the invariance of the scalar $\tilde{X}^\top\Sigma^{-1}\boldsymbol{\epsilon} \in R$ to the transpose operation, as well as the symmetry of the covariance matrix (and its inverse). □

Proposition E.2 (Expected value of uniform-degradation weighted-mean)**.** *Let X be a T-long episodic signal of length $n = KT$ for some $K \in \mathbb{N}$ (i.e., integer number of episodes), with parameters $\boldsymbol{\mu_0}, \Sigma_0$. The expected value of $\frac{1}{n}s_{unif}(X|\Sigma_0)$ defined in Definition D.6 is $\frac{1}{T}(W_0 \cdot \boldsymbol{\mu_0})$.*

Proof. Using the T-periodicity of W and $\boldsymbol{\mu}$ (see Eq. (4)), we have $E[\frac{1}{n}s_{unif}] = E[\frac{1}{n}W \cdot X] = \frac{1}{n}W \cdot \boldsymbol{\mu}(\boldsymbol{\mu_0}) = \frac{1}{n}K \cdot (W_0 \cdot \boldsymbol{\mu_0}) = \frac{1}{T}(W_0 \cdot \boldsymbol{\mu_0})$. □

Proposition E.3 (Consistency of uniform-degradation weighted-mean)**.** *$\frac{1}{n}s_{unif}$ defined in Definition D.6 is consistent with relation to the expected value $\frac{1}{T}(W_0 \cdot \boldsymbol{\mu_0})$ calculated in Proposition E.2, i.e., $\forall\epsilon > 0$: $lim_{n\to\infty} P\left(|\frac{1}{n}s_{unif} - \frac{1}{T}(W_0 \cdot \boldsymbol{\mu_0})| \geq \epsilon\right) = 0$.*

Proof. The consistency is proven through the L.L.N over the i.i.d episodes, where the last possibly-partial episode becomes negligible in the limit of infinitely-many episodes.

Using the episodic index decomposition of Definition C.1, and the notations of W_0, W_τ from Definition D.6, we can write

$$\begin{aligned} \frac{1}{n}s_{unif} - \frac{1}{T}(W_0 \cdot \boldsymbol{\mu_0}) &= \frac{1}{n}WX - \frac{1}{T}(W_0 \cdot \boldsymbol{\mu_0}) = \frac{1}{n}\sum_{t=1}^n [w_t X_t] - \frac{1}{T}(W_0 \cdot \boldsymbol{\mu_0}) \\ &= \left[\frac{1}{n}\sum_{k=0}^{K-1}\sum_{\tau=1}^{T}(W_0)_\tau X_{kT+\tau} - \frac{KT}{n}W_0\boldsymbol{\mu_0}\right] + \left[\frac{1}{n}\sum_{\tau=1}^{\tau_0}(W_{\tau_0})_\tau X_{kT+\tau} - \frac{\tau_0}{n}W_0\boldsymbol{\mu_0}\right] \\ &= \frac{1}{n}\sum_{k=0}^{K-1}\sum_{\tau=1}^{T}(W_0)_\tau(X_{kT+\tau} - (\boldsymbol{\mu_0})_\tau) + \frac{1}{n}\sum_{\tau=1}^{\tau_0}(W_{\tau_0})_\tau(X_{kT+\tau} - (\boldsymbol{\mu_0})_\tau) \\ &= \frac{1}{n}\sum_{k=0}^{K-1} S_k + \frac{1}{n}S_{K,\tau_0}^{tail} \end{aligned}$$

where $\tau_0 := \tau(n,T)$, $S_k := \sum_{\tau=1}^{T}(W_0)_\tau(X_{kT+\tau} - (\boldsymbol{\mu_0})_\tau)$ and $S_{K,\tau_0}^{tail} := \sum_{\tau=1}^{\tau_0}(W_{\tau_0})_\tau(X_{kT+\tau} - (\boldsymbol{\mu_0})_\tau)$.

To prove consistency we have to show that $\forall \epsilon > 0 : \lim_{n\to\infty} P\left(|\frac{1}{n}s_{unif} - \frac{1}{T}(W_0 \cdot \boldsymbol{\mu_0})| \geq \epsilon\right) = 0$. Indeed, given $\epsilon > 0$, we have

$$\begin{aligned}
&\lim_{n\to\infty} P\left(\left|\frac{1}{n}s_{unif} - \frac{1}{T}(W_0 \cdot \boldsymbol{\mu_0})\right| \geq \epsilon\right) \\
&\leq \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=0}^{K-1} S_k\right| \geq \epsilon/2 \vee \left|\frac{1}{n}S^{tail}_{K,\tau_0}\right| \geq \epsilon/2\right) \\
&\leq \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=0}^{K-1} S_k\right| \geq \epsilon/2\right) + P\left(\left|\frac{1}{n}S^{tail}_{K,\tau_0}\right| \geq \epsilon/2\right)
\end{aligned}$$

where $P\left(\left|\frac{1}{n}\sum_{k=0}^{K-1} S_k\right| \geq \epsilon/2\right) \to 0$ according to the Law of Large Numbers applied to the i.i.d sequence $\{S_k\}$; and

$$\begin{aligned}
&\lim_{n\to\infty} P\left(\left|\frac{1}{n}S^{tail}_{K,\tau_0}\right| \geq \epsilon/2\right) \\
&\qquad \leq \lim_{n\to\infty} P\left(\sum_{\tau=1}^{T} |\max_{1\leq\tilde{\tau}\leq T}(W_{\tilde{\tau}})_\tau| \cdot |X_{kT+\tau} - (\boldsymbol{\mu_0})_\tau| \geq \frac{n\epsilon}{2}\right) \\
&\qquad = \lim_{n\to\infty} P\left(\sum_{\tau=1}^{T} |\max_{1\leq\tilde{\tau}\leq T}(W_{\tilde{\tau}})_\tau| \cdot |X_{\tau} - (\boldsymbol{\mu_0})_\tau| \geq \frac{n\epsilon}{2}\right) = 0
\end{aligned}$$

□

Lemma 1 (Maximum likelihood with relation to the complex hypothesis of uniform-degradation)**.** *Under the setup of Theorem D.1, denote* $s_0 \coloneqq W \cdot \boldsymbol{\mu} - \epsilon_0[\mathbf{1}^\top\Sigma^{-1}\mathbf{1}]$*.* $\lambda_{LR}(H_0, H_A^{unif}(\epsilon_0)|\{X_t\}_{t=1}^n)$ *is minimized by* $\epsilon = \epsilon_0$ *if* $s_{unif} \geq s_0$*, and by* $\epsilon = \frac{W\cdot\boldsymbol{\mu} - s_{unif}}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}}$ *if* $s_{unif} \leq s_0$*.*

Proof. Applying Proposition E.1 to Definition D.4 yields

$$\begin{aligned}
&\lambda_{LR}(H_0, H_A^{unif}(\epsilon_0)|\{X_t\}_{t=1}^n) = \\
&\quad \min_{\epsilon\geq\epsilon_0} 2\epsilon[W\tilde{X}] + \epsilon^2[\mathbf{1}^\top\Sigma^{-1}\mathbf{1}] = \min_{\epsilon\geq\epsilon_0} P(\epsilon)
\end{aligned} \tag{8}$$

where $P(\epsilon)$ is a parabola with respect to ϵ, with leading coefficient $\mathbf{1}^\top\Sigma^{-1}\mathbf{1} > 0$ (since the full-rank covariance matrix Σ is necessarily positive definite) and minimum $\epsilon_{min} = -\frac{2W\tilde{X}}{2[\mathbf{1}^\top\Sigma^{-1}\mathbf{1}]} = \frac{W\boldsymbol{\mu} - s_{unif}}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}}$ (remember that $\tilde{X} = X - \boldsymbol{\mu}$). If $s_{unif} \leq s_0$ then $\epsilon_{min} \geq \epsilon_0$ and $\min_{\epsilon\geq\epsilon_0} P(\epsilon)$ is minimized by $\epsilon = \epsilon_{min}$. If $s_{unif} \geq s_0$ then $\epsilon_{min} \leq \epsilon_0$ (i.e., ϵ_0 is to the right of the minimum of the parabola), hence $\forall \epsilon > \epsilon_0 : P(\epsilon) > P(\epsilon_0)$, and $\min_{\epsilon\geq\epsilon_0} P(\epsilon)$ is minimized by $\epsilon = \epsilon_0$. □

Proof of Theorem D.1 (also compactly formulated in Theorem 4.1): Given $\alpha \in (0, 1)$, consider a $\tilde{\kappa}$-threshold-test (Definition D.7) with relation to the log-likelihood $\lambda_{LR}(H_0, H_A^{unif}(\epsilon_0)|\{X_t\}_{t=1}^n)$ (and with edge-case rejection-probability $\rho \in [0, 1]$), such that the significance level of the test is $1 - \alpha$. Since $\lambda_{LR} = 2\ln(LR)$ is monotonously increasing with relation to the likelihood-ratio, then according to Neyman-Pearson lemma (Neyman et al., 1933) this test has the greatest power among all tests with significance $\tilde{\alpha} \leq \alpha$. We will show that this test is equivalent to a threshold-test on the uniform-degradation weighted-mean.

According to Lemma 1, we have

$$\begin{aligned}
&\lambda_{LR}(H_0, H_A^{unif}(\epsilon_0)|\{X_t\}_{t=1}^n) \\
&\quad = \min_{\epsilon\geq\epsilon_0} 2\epsilon[W\tilde{X}] + \epsilon^2[\mathbf{1}^\top\Sigma^{-1}\mathbf{1}] \\
&\quad = \begin{cases} 2\epsilon_0[W\tilde{X}] + \epsilon_0^2[\mathbf{1}^\top\Sigma^{-1}\mathbf{1}] & \text{if } s_{unif} \geq s_0 \\ 2\frac{W\cdot\boldsymbol{\mu} - s_{unif}}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}}[W\tilde{X}] + [\frac{W\cdot\boldsymbol{\mu} - s_{unif}}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}}]^2[\mathbf{1}^\top\Sigma^{-1}\mathbf{1}] & \text{if } s_{unif} \leq s_0 \end{cases} \\
&\quad = \begin{cases} 2\epsilon_0 s_{unif} - 2\epsilon_0 W\boldsymbol{\mu} + \epsilon_0^2[\mathbf{1}^\top\Sigma^{-1}\mathbf{1}] & \text{if } s_{unif} \geq s_0 \\ -\frac{(s_{unif} - W\boldsymbol{\mu})^2}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}} & \text{if } s_{unif} \leq s_0 \end{cases}
\end{aligned} \tag{9}$$

Clearly λ_{LR} is strictly increasing with s_{unif} in $(-\infty, s_0]$. Note that in the case $s_{unif} \geq s_0$, λ_{LR} is the parabola $P(s_{unif}) = -(s_{unif} - W\boldsymbol{\mu})^2$ (up to a positive multiplicative factor), whose maximum is $s_{max} = W\boldsymbol{\mu}$. Since in this case $s_{unif} \geq s_0 = W\boldsymbol{\mu} - \epsilon_0[\mathbf{1}^\top\Sigma^{-1}\mathbf{1}] \leq W\boldsymbol{\mu} = s_{max}$, then s_{unif} is to the left of the maximum of the parabola, and hence λ_{LR} is strictly increasing with s_{unif} in $[s_0, \infty)$.

Since λ_{LR} is strictly increasing with s_{unif} in both $(-\infty, s_0]$ and $[s_0, \infty)$, then it is strictly monotonously increasing with s_{unif} in $\mathbb{R}$. Hence there exists $\kappa \in \mathbb{R}$ such that $\lambda_{LR} < \tilde{\kappa} \Leftrightarrow s_{unif} < \kappa$, and the tests are equivalent. □

Lemma 2 (Properties of statistics under uniform-degradation)**.** *Let X be a T-long episodic signal of length $n = KT$ for some $K \in \mathbb{N}$ (i.e., integer number of episodes), with parameters $\boldsymbol{\mu_0} - \epsilon \cdot \mathbf{1} \in \mathbb{R}^T, \Sigma_0 \in \mathbb{R}^{T\times T}$. Denote $s_{simp} = \sum_{t=1}^n X_t$ as in Eq.* (6) *and $s_{unif} = WX$ as in Definition D.6. Then we have:*

$$
\begin{aligned}
E[s_{simp}] &= K\mathbf{1}^\top\boldsymbol{\mu_0} - KT\epsilon \\
E[s_{unif}] &= KW_0\boldsymbol{\mu_0} - \epsilon KW_0\mathbf{1} \\
\mathit{Var}(s_{simp}) &= K\mathbf{1}^\top\Sigma_0\mathbf{1} \\
\mathit{Var}(s_{unif}) &= K\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1}
\end{aligned}
$$

Proof. The first 3 identities are straight-forward:

$$
\begin{aligned}
E[s_{simp}] &= \sum_{k=0}^{K-1}\sum_{\tau=1}^{T}(\boldsymbol{\mu_0})_\tau - \epsilon = K\mathbf{1}^\top\boldsymbol{\mu_0} - KT\epsilon \\
E[s_{unif}] &= \sum_{k=0}^{K-1} W_0 \cdot (\boldsymbol{\mu_0} - \epsilon\mathbf{1}) = KW_0\boldsymbol{\mu_0} - \epsilon KW_0\mathbf{1} \\
\mathrm{Var}(s_{simp}) &= \sum_{i,j=1}^{n} \mathrm{Cov}(X_i, X_j) = K\sum_{i,j=1}^{T}\mathrm{Cov}(X_i, X_j) = K\mathbf{1}^\top\Sigma_0\mathbf{1}
\end{aligned}
$$

For the last identity denote $Y \coloneqq \Sigma_0^{-1}X \in \mathbb{R}^n$ (i.e., $Y_i = \sum_{m=1}^T(\Sigma_0^{-1})_{im}X_m$), such that $s_{unif} = \sum_i Y_i$.

$$
\begin{aligned}
\mathrm{Var}(s_{unif}) &= \sum_{i,j=1}^{n}\mathrm{Cov}(Y_i, Y_j) \\
&= \sum_{i,j=1}^{n}\mathrm{Cov}(\sum_{m=1}^{n}\Sigma_{im}^{-1}X_m, \sum_{l=1}^{n}\Sigma_{jl}^{-1}X_l) \\
&= \sum_{i,j,m,l=1}^{n}\Sigma_{im}^{-1}\Sigma_{jl}^{-1}\mathrm{Cov}(X_m, X_l) \\
&= K\sum_{i,j,m,l=1}^{T}(\Sigma_0)_{im}^{-1}(\Sigma_0)_{jl}^{-1}\mathrm{Cov}(X_m, X_l) \\
&= K\sum_{i,j,m,l=1}^{T}(\Sigma_0)_{jl}^{-1}(\Sigma_0)_{im}^{-1}(\Sigma_0)_{ml} \\
&= K\sum_{i,j,l=1}^{T}(\Sigma_0)_{jl}^{-1}\left((\Sigma_0)_{i\cdot}^{-1}\cdot(\Sigma_0)_{\cdot l}\right) \\
&= K\sum_{i,j,l=1}^{T}(\Sigma_0)_{jl}^{-1}\delta_{il} = K\sum_{i,j=1}^{T}(\Sigma_0)_{ji}^{-1} = K\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1}
\end{aligned}
$$

□

Proof of Proposition D.1: Both $\sqrt{K}\tilde{s}^K_{simp}$ and $\sqrt{K}\tilde{s}^K_{unif}$ defined in Eq. (6) are under H_0 the sums of K i.i.d variables with mean 0 and variance 1. Thus, according to the Central Limit Theorem (Petrov, 1972; Irwin, 2006), both converge-in-distribution to the standard normal distribution under H_0:

$$\tilde{s}^K_{simp} \xrightarrow[K\to\infty]{D} N(0,1)$$
$$\tilde{s}^K_{unif} \xrightarrow[K\to\infty]{D} N(0,1)$$

Hence, from the definition of convergence in distribution, we have

$$\lim_{K\to\infty} P\left(\tilde{s}^K_{simp} \le q^0_\alpha \middle| H_0\right) = \lim_{K\to\infty} F_{\tilde{s}^K_{simp}|H_0}\left(q^0_\alpha\right) = \Phi\left(q^0_\alpha\right) = \alpha$$
$$\lim_{K\to\infty} P\left(\tilde{s}^K_{unif} \le q^0_\alpha \middle| H_0\right) = \lim_{K\to\infty} F_{\tilde{s}^K_{unif}|H_0}\left(q^0_\alpha\right) = \Phi\left(q^0_\alpha\right) = \alpha$$

where $F_{s|H}$ is the Cumulative Distribution Function of the random variable s under the hypothesis H, and Φ is of the standard normal distribution.

Note that $\tilde{s}^K_{simp}, \tilde{s}^K_{unif}$ can be computed from s_{simp}, s_{unif} by substituting Lemma 2 (with $\epsilon = 0$, corresponding to H_0) in Eq. (6):

$$\begin{aligned} \tilde{s}^K_{simp} &= \frac{s_{simp} - K\mathbf{1}^\top \boldsymbol{\mu_0}}{\sqrt{K\mathbf{1}^\top \Sigma_0 \mathbf{1}}} \\ \tilde{s}^K_{unif} &= \frac{s_{unif} - KW_0\boldsymbol{\mu_0}}{\sqrt{K\mathbf{1}^\top \Sigma_0^{-1} \mathbf{1}}} \end{aligned} \tag{10}$$

and by substituting Lemma 2 with $\epsilon > 0$ in Eq. (10), we also have the properties of $\tilde{s}^K_{simp}, \tilde{s}^K_{unif}$ under H^ϵ_A:

$$E\left[\tilde{s}^K_{simp}|H^\epsilon_A\right] = -\frac{KT\epsilon}{\sqrt{K\mathbf{1}^\top\Sigma_0\mathbf{1}}} = -\frac{\sqrt{K}T\epsilon}{\sqrt{\mathbf{1}^\top\Sigma_0\mathbf{1}}}$$
$$E\left[\tilde{s}^K_{unif}|H^\epsilon_A\right] = -\frac{KW_0\mathbf{1}\epsilon}{\sqrt{K\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1}}} = -\frac{\sqrt{K}W_0\mathbf{1}\epsilon}{\sqrt{\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1}}}$$
$$\text{Var}(\tilde{s}^K_{simp}|H^\epsilon_A) = \text{Var}(\tilde{s}^K_{unif}|H^\epsilon_A) = 1$$

Accordingly, using the Central Limit Theorem again, we have under H^ϵ_A:

$$\tilde{s}^K_{simp} + \frac{\sqrt{K}T\epsilon}{\sqrt{\mathbf{1}^\top\Sigma_0\mathbf{1}}} \xrightarrow[K\to\infty]{D} N(0,1)$$
$$\tilde{s}^K_{unif} + \frac{\sqrt{K}W_0\mathbf{1}\epsilon}{\sqrt{\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1}}} \xrightarrow[K\to\infty]{D} N(0,1)$$

and by the definition of convergence in distribution, we receive

$$\begin{aligned} &\lim_{K\to\infty} P\left(\tilde{s}^K_{simp} \le q^0_\alpha \middle| H^\epsilon_A\right) = \lim_{K\to\infty} F_{\tilde{s}^K_{simp}|H^\epsilon_A}\left(q^0_\alpha\right) = \\ &\quad \lim_{K\to\infty} F_{\tilde{s}^K_{simp} + \frac{\sqrt{K}T\epsilon}{\sqrt{\mathbf{1}^\top\Sigma_0\mathbf{1}}}|H_0}\left(q^0_\alpha\right) = \lim_{K\to\infty}\Phi\left(q^0_\alpha + \frac{\sqrt{K}T\epsilon}{\sqrt{\mathbf{1}^\top\Sigma_0\mathbf{1}}}\right) = 1 \\ &\lim_{K\to\infty} P\left(\tilde{s}^K_{unif} \le q^0_\alpha \middle| H^\epsilon_A\right) = \lim_{K\to\infty} F_{\tilde{s}^K_{unif}|H^\epsilon_A}\left(q^0_\alpha\right) = \\ &\quad \lim_{K\to\infty} F_{\tilde{s}^K_{unif} + \frac{\sqrt{K}W_0\mathbf{1}\epsilon}{\sqrt{\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1}}}|H_0}\left(q^0_\alpha\right) = \lim_{K\to\infty}\Phi\left(q^0_\alpha + \frac{\sqrt{K}W_0\mathbf{1}\epsilon}{\sqrt{\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1}}}\right) = 1 \end{aligned} \tag{11}$$

□

Proof of Theorem D.2 (also compactly formulated in Theorem 4.2): Following identical reasoning to the proof of Proposition D.1 with ϵ replaced by $\epsilon/\sqrt{K}$, and recalling that $W_0 = \mathbf{1}^\top \Sigma_0^{-1}$ (Definition D.6), we receive the analog of Eq. (11):

$$\begin{aligned}
\lim_{K\to\infty} P\left(\tilde{s}^K_{simp} \leq q^0_\alpha \middle| H^{\epsilon,K}_A\right) &= \Phi\left(q^0_\alpha + \frac{T\epsilon}{\sqrt{\mathbf{1}^\top \Sigma_0 \mathbf{1}}}\right) \\
\lim_{K\to\infty} P\left(\tilde{s}^K_{unif} \leq q^0_\alpha \middle| H^{\epsilon,K}_A\right) &= \Phi\left(q^0_\alpha + \epsilon\sqrt{\mathbf{1}^\top \Sigma_0^{-1} \mathbf{1}}\right)
\end{aligned} \tag{12}$$

To complete the proof, since Φ is monotonously increasing, we only have to show that $\frac{T}{\sqrt{\mathbf{1}^\top\Sigma_0\mathbf{1}}} \leq \sqrt{\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1}}$, or equivalently $\frac{T}{\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1}} \leq \frac{\mathbf{1}^\top\Sigma_0\mathbf{1}}{T}$, which can be seen as a matrix-form generalization for the harmonic-algebraic means inequality.

Since the invertible covariance matrix Σ_0 is necessarily symmetric and positive definite, it has a symmetric positive definite square-root $R^2 = \Sigma_0$. Since $\mathbf{1}^\top\Sigma_0\mathbf{1} = \mathbf{1}^\top R^\top R\mathbf{1} = \|R\mathbf{1}\|^2$ and $\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1} = \|R^{-1}\mathbf{1}\|^2$, we indeed have by Cauchy-Schwarz inequality

$$(\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1})(\mathbf{1}^\top\Sigma_0\mathbf{1}) = \|R^{-1}\mathbf{1}\|^2 \cdot \|R\mathbf{1}\|^2 \geq ((\mathbf{1}^\top R^{-1})(R\mathbf{1}))^2 = (\mathbf{1}^\top\mathbf{1})^2 = T^2 \tag{13}$$

□

Proof of Proposition 4.1 Since Σ_0 is symmetric it is orthogonally diagonalizable, i.e., $\Sigma_0 = U^\top AU$ where A is diagonal and U is orthogonal. Since Σ_0 is positive-definite (note that Definition D.1 assumes full-rank covariance matrix), its eigenvalues are positive, i.e., $\forall 1 \leq i \leq T : \lambda_i = A_{ii} > 0$. We also have $\mathbf{1}^\top\Sigma_0\mathbf{1} = \mathbf{1}^\top U^\top AU\mathbf{1} = u^\top Au$ (where $u = U\mathbf{1}$), and $\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1} = \mathbf{1}^\top U^\top A^{-1}U\mathbf{1} = u^\top A^{-1}u$.

From this we receive

$$\begin{aligned}
G^2 =& \frac{(\mathbf{1}^\top\Sigma_0^{-1}\mathbf{1})(\mathbf{1}^\top\Sigma_0\mathbf{1})}{T^2} = \frac{(u^\top A^{-1}u)(u^\top Au)}{T^2} \\
=& \frac{1}{T^2}(\sum_{i=1}^T u_i^2\lambda_i)(\sum_{j=1}^T u_j^2/\lambda_j) = \frac{1}{T^2}\sum_{i,j=1}^T (u_iu_j)^2\frac{\lambda_i}{\lambda_j} \\
=& \frac{1}{2T^2}\sum_{i,j=1}^T (u_iu_j)^2\left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i}\right) \\
=& \frac{1}{2T^2}\sum_{i,j=1}^T (u_iu_j)^2\frac{\lambda_i^2 + \lambda_j^2}{\lambda_i\lambda_j} \\
=& \frac{1}{2T^2}\sum_{i,j=1}^T (u_iu_j)^2\frac{(\lambda_i - \lambda_j)^2 + 2\lambda_i\lambda_j}{\lambda_i\lambda_j} \\
=& \frac{1}{2T^2}\left[2\sum_{i,j=1}^T (u_iu_j)^2 + \sum_{i,j=1}^T (u_iu_j)^2\frac{(\lambda_i - \lambda_j)^2}{\lambda_i\lambda_j}\right] \\
=& \frac{1}{2T^2}\left[2u^\top u + \sum_{i,j=1}^T (u_iu_j)^2\frac{(\lambda_i - \lambda_j)^2}{\lambda_i\lambda_j}\right] \\
=& \frac{1}{2T^2}\left[2(\mathbf{1}^\top I\mathbf{1}) + \sum_{i,j=1}^T (u_iu_j)^2\frac{(\lambda_i - \lambda_j)^2}{\lambda_i\lambda_j}\right] \\
=& 1 + \frac{1}{2T^2}\sum_{i,j=1}^T (u_iu_j)^2\frac{(\lambda_i - \lambda_j)^2}{\lambda_i\lambda_j}
\end{aligned}$$

and since $\forall i: u_i = U_i.\mathbf{1} \neq 0$ as the sum of a row of an orthogonal matrix, we only need to denote $w_{ij} := \frac{(u_i u_j)^2}{2T^2 \lambda_i \lambda_j} > 0$. $\square$

Lemma 3 (Sensitivity of the minimum to deviations in the elements)**.** *Let a finite set $\mathbb{A}$, functions $f, g : \mathbb{A} \to \mathbb{R}$, and $\epsilon > 0$. Note that since $\mathbb{A}$ is finite, both f, g are bounded and denote $|g| \leq G$ an upper bound. Denote $y := min_{a \in \mathbb{A}} f(a) + \epsilon g(a)$ and $\tilde{y} := min_{a \in \mathbb{A}} f(a)$. Then $|y - \tilde{y}| \leq 3\epsilon G$.*

Proof. Denote by $a_0, \tilde{a}_0$ the minimizers of $y, \tilde{y}$ respectively, i.e., $y = f(a_0) + \epsilon g(a_0) \leq f(\tilde{a}_0) + \epsilon g(\tilde{a}_0)$ and $\tilde{y} = f(\tilde{a}_0) \leq f(a_0)$. From the last two inequalities we get $0 \leq f(a_0) - f(\tilde{(a)}_0) \leq \epsilon(g(\tilde{a}_0) - g(a_0))$. Finally from the triangle inequality,

$$
\begin{aligned}
|y - \tilde{y}| &= |f(a_0) - f(\tilde{a}_0) + \epsilon g(a_0)| \leq |\epsilon(g(\tilde{a}_0) - g(a_0))| + |\epsilon g(a_0)| \\
&\leq \epsilon \left[|g(\tilde{a}_0)| + |g(a_0))| + |g(a_0)| \right] \leq 3\epsilon G
\end{aligned}
$$

$\square$

Proof of Theorem 4.3: Similarly to the proof of Theorem D.1, we wish to show that the log-likelihood-ratio λ_{LR} from Lemma 1 – after substituting Definition D.10 – is strictly monotonously increasing with s_{part}.

Given $a_0 \in \{0,1\}^T$, let $a(a_0) \in \{0,1\}^n$ be its T-periodic completion to n dimensions, and recall the notations $m(p) = \lceil pT \rceil, n = KT + \tau_0$. Then we have

$$
\begin{aligned}
&\lambda_{LR}(H_0, H_A^{part} | \{X_t\}_{t=1}^n) = \\
&\quad = \min_{a_0 \in A_m^T} 2\epsilon a^\top \Sigma^{-1} \tilde{X} + \epsilon^2 a^\top \Sigma^{-1} a \\
&\quad = 2\epsilon \cdot \min_{a_0 \in A_m^T} \left(a^\top \Sigma^{-1} \tilde{X} + 0.5\epsilon (a^\top \Sigma^{-1} a) \right)
\end{aligned}
$$

Denote $f(a_0) = a(a_0)^\top \Sigma^{-1} \tilde{X}$, $g(a_0) = 0.5 a(a_0)^\top \Sigma^{-1} a(a_0)$ and $y = \min_{a_0 \in A_m^T} f(a_0) + \epsilon g(a_0)$, such that $\lambda_{LR} = 2\epsilon y$ is monotonously increasing with y. Note that

$$
\begin{aligned}
\min_{a_0 \in A_m^T} f(a_0) &= \min_{a_0 \in A_m^T} a^\top \Sigma^{-1} \tilde{X} \\
&= \min_{a_0 \in A_m^T} \sum_{k=0}^{K-1} \sum_{\substack{\tau=1 \\ (a_0)_\tau = 1}}^{T} \tilde{s}_{kT+\tau} + \sum_{\substack{\tau=1 \\ (a_0)_\tau = 1}}^{\tau_0} \tilde{s}_{kT+\tau} \\
&= \min_{a_0 \in A_m^T} \sum_{\tau \in o(a_0)} \tilde{S}_\tau \\
&= s_{part}(X)
\end{aligned}
$$

Also note that the term $g(a_0)$ is bounded:

$$
\forall a_0 \in A_m^T : |g(a_0)| \leq \frac{1}{2} \sum_{i,j=1}^{T} |(\Sigma^{-1})_{ij}| \leq \frac{K+1}{2} \sum_{i,j=1}^{T} |(\Sigma_0^{-1})_{ij}|
$$

Hence, by Lemma 3, we have

$$
|y - s_{part}| \leq \epsilon \frac{3(K+1)}{2} \sum_{i,j=1}^{T} |(\Sigma_0^{-1})_{ij}| = \mathcal{O}(\epsilon) \tag{14}
$$

(where $\mathcal{O}(\epsilon)$ is defined with relation to $\epsilon \to 0$).

Now consider the α-quantile of y under H_0, denoted $\tilde{\kappa} = q_\alpha(y|H_0)$. By construction $P(y \leq \tilde{\kappa}|H_0) = \alpha$ (up to non-continuous probability density in the edge case $y = \tilde{\kappa}$). According to Neyman-Pearson lemma, a threshold-test on y has the greatest power P_α among all statistical tests with significance level $\leq \alpha$ (see the proof of Theorem D.1 for more details), i.e., $P_\alpha = P(y \leq \tilde{\kappa}|H_A^{part}) = F_{y|H_A^{part}}(\tilde{\kappa})$ (where $F_{s|H}$ is the Cumulative Distribution Function of the variable s given the hypothesis H).

Denote the α-quantile of the actual test-statistic s_{part} by $\kappa = q_\alpha(s_{part}|H_0)$. Since $\forall X \in \mathbb{R}^n : |y - s_{part}| = \mathcal{O}(\epsilon)$ (Eq. (14)), we also have $\left|\tilde{\kappa} - \kappa\right| = \left|q_\alpha(y|H_0) - q_\alpha(s_{part}|H_0)\right| = \mathcal{O}(\epsilon)$. Hence, along with Eq. (14), we have

$$\begin{aligned} P\left(s_{part} \le \kappa | H_A^{part}\right) &\ge P\left(y + \mathcal{O}(\epsilon) \le \tilde{\kappa} - \mathcal{O}(\epsilon) \middle| H_A^{part}\right) \\ &= P\left(y \le \tilde{\kappa} - \mathcal{O}(\epsilon) \middle| H_A^{part}\right) \\ &= F_{y|H_A^{part}}\left(\tilde{\kappa} - \mathcal{O}(\epsilon)\right) \\ &= F_{y|H_A^{part}}\left(\tilde{\kappa}\right) - \mathcal{O}(\epsilon) \\ &= P_\alpha - \mathcal{O}(\epsilon) \end{aligned}$$

where the second-to-last equality is true since $F_{y|H_A^{part}}(x)$ has a finite derivative at $x = \tilde{\kappa}$, as the CDF of the minimum of the normal variables $\{a^\top \Sigma^{-1} \tilde{X} + 0.5\epsilon(a^\top \Sigma^{-1} a)\}_{a \in A_m^T}$. $\square$

F. Bootstrap for Sequential Tests: Extended Discussion

Section 5 describes a mechanism for sequential hypothesis testing with relation to episodic signals. The mechanism simply runs individual hypothesis tests repeatedly with a constant significance level α, similarly to the concept of α-spending functions (Lan, 1994; PennState College of Science), and in particular to Pocock approach (Pocock, 1977).

Note that Pocock's constant α-spending function is often avoided, as it is claimed to spend "too much" α-budget in the beginning of the sequential test on account of its end. In our online setup this time-homogeneous approach is welcome, as we do not to rely on well-defined beginning and end. However, in contrast to Pocock, we cannot assume independence between nor normality of the aggregative parts of the data.

The sequential test (described in Algorithm 5) uses a constant manually-determined lookback-horizon h. Any individual test at time $t = kT + \tau$ runs the simple threshold-test of Algorithm 3 on the signal $X_{(k-h)T}, ..., X_{kT+\tau}$, i.e., it looks exactly $h + \tau/T$ episodes back. In practice, n_h different lookback-horizons $h_1, ..., h_{n_h}$ can be used simultaneously, such that at any point of time, we reject H_0 if any of the lookback tests rejects it. This allows us to detect slight changes which are only detectable over large amounts of data (large h); while still allowing quick detection of larger abrupt changes, without mixing them with older irrelevant data (small h).

In order to determine the significance level α for the individual tests within the sequential test, we use the bootstrap mechanism described in Algorithm 1 (also see extended pseudo-code in Algorithm 4 in Appendix G). The mechanism simulates sequential tests using bootstrap-sampling of $\max(h_1, ..., h_{n_h}) + \tilde{h}$ episodes (length of simulation + maximum lookback horizon) from a reference dataset of N episodes assumed to be i.i.d. Once the episodes are sampled, the simulation runs $\tilde{h}$ episodes without stopping condition, keeps track of the resulted p-values, and eventually returns the minimal p-value among all the individual tests. This simulative process is repeated $\tilde{B}$ times with different bootstrap-samples, and the α_0-quantile among all the minimal-p-values is chosen as the individual-test significance level α. Indeed, α_0 is the relative part of bootstrap-samples in which at least one individual test returned p-value smaller than α.

The sequential bootstrap mechanism of Algorithm 1 may look computationally overwhelming since it runs a bootstrap that calls another bootstrap (Algorithm 2, called through Algorithm 3). However, if the sequential test runs individual tests in n_h different lookback-horizons (where typically $n_h \le 3$) in frequency of F test-points per episode, then the inner bootstrap of Algorithm 2 will only be called $n_h \cdot F$ times in total (thanks to the bootstrap-storage mechanism described in Algorithm 3). Also note that in spite of its name, the whole sequential bootstrap algorithm is intended to run only once (and not sequentially) – after the reference dataset becomes available.

As a practical remark for implementation, note that Algorithm 3 necessarily returns p-value$\ge \frac{1}{B+1}$, which is the resolution of the inner bootstrap. Now consider the case where in Algorithm 1, in more than α_0 of the simulated sequential tests, there is certain individual test whose return value is $\frac{1}{B+1}$. In other words, the bootstrap found that under H_0, with probability higher than α_0, a sequential test of length $\tilde{h}$ will encounter the most extreme possible result of Algorithm 3 at least once. In that case Algorithm 5 would not be able to distinguish any degradation from H_0: we would have the individual test threshold set to $\alpha = \text{quantile}_{\alpha_0}(P) = \frac{1}{B+1}$, which can never be overcome. For this reason, Algorithm 5 makes sure to check whether $\alpha = \frac{1}{B+1}$. Possible solutions in this situation are increase of B for better resolution, or reduction of the required significance level through either $\tilde{h}$ or α_0.

Multiple test-statistics: Every iteration, Algorithm 5 calculates the test-statistic for multiple lookback horizons, where Algorithm 1 is responsible of controlling the family-wise type-I error rate through the test-thresholds. In a similar manner, Algorithm 5 can be generalized to run multiple test-statistics in parallel: simply iterate over the statistics the same as iterating over the lookback-horizons.

Heterogeneous test-statistics should provide more robustness to the alternative hypothesis, since every statistic is often affected differently by every alternative hypothesis. This comes at the cost of reduced sensitivity of each statistic, expressed through decrease of the test-thresholds by Algorithm 1.

G. Algorithms (Pseudocode)

This appendix concentrates the pseudo-code of the algorithms for hypothesis testing and for bootstrap-based α tuning, in the contexts of both individual and sequential tests. Algorithm 4 is a more detailed version of the pseudo-code of Algorithm 1.

Algorithm 2: Individual_test_bootstrap

Input: $x \in \mathbb{R}^{N \times T}$ assumed to be drawn from a T-long episodic signal; sample size $n = KT + \tau_0 \in \mathbb{N}$; a test-statistic function $s : \mathbb{R}^n \to \mathbb{R}$; number of repetitions $B \in \mathbb{N}$;
Output: test-statistic bootstrap distribution $S \in \mathbb{R}^B$;
Algorithm:
Initialize $S \in \mathbb{R}^B$;
for *b in 1:B* **do**
    // sample
    Initialize $y \in \mathbb{R}^n$;
    for *k in 0:K-1* **do**
        Sample j uniformly from $(1, ..., N)$;
        $y[kT + 1 : kT + T] \leftarrow (x_{j1}, ..., x_{jT})$;
    Sample j uniformly from $(1, ..., N)$;
    $y[KT + 1 : KT + \tau_0] \leftarrow (x_{j1}, ..., x_{j\tau_0})$;
    // calculate
    $S_b \leftarrow s(y)$;
Return S;

Algorithm 3: Individual_degradation_test

Input: reference episodic signal $x_0 \in \mathbb{R}^{N \times T}$; test data $x \in \mathbb{R}^n$; a test-statistic function $s : \mathbb{R}^n \to \mathbb{R}$; bootstrap repetitions $B \in \mathbb{N}$; bootstrap distributions storage BS; allowed type-I error rate $\alpha \in (0, 1)$;
Output: rejection $\in \{0, 1\}$; P-value $p \in \mathbb{R}$;
Algorithm:
if $BS[n]$ *not exists* **then**
    $BS[n] \leftarrow$ Individual_test_bootstrap(T, N, x_0, n, s, B);   (Algorithm 2)
$S \leftarrow BS[n]$;
$\kappa_\alpha \leftarrow \text{quantile}_\alpha(S)$;
$y \leftarrow s(x)$;
$count \leftarrow |\{b \in \{1, ..., B\} | S_b \leq y\}|$;
$p \leftarrow \frac{1+\text{count}}{1+B}$;
$reject = 1$ if $y < \kappa_\alpha$ else 0;   (or equivalently, 1 if $p < \alpha$ else 0)
Return $reject$, p;

Algorithm 4: Sequential_test_bootstrap

Input: $x \in \mathbb{R}^{N \times T}$ assumed to be drawn from a T-long episodic signal; test-statistic function s; inner-bootstrap repetitions $B \in \mathbb{N}$; inner-bootstrap storage BS; tests frequency $d \in [1, T]$ and lookback horizons $h_1, ..., h_{n_h} \in \mathbb{N}$; sequential test length $\tilde{h} \in \mathbb{N}$; outer-bootstrap repetitions $\tilde{B} \in \mathbb{N}$;

Output: bootstrap-distribution $P \in [0,1]^{\tilde{B}}$ of the minimal-p-value in a sequential test of $\tilde{h}$ episodes under H_0;

Algorithm:

```
Initialize P = (1, ..., 1) ∈ [0,1]^B̃;
h_max ← max(h_1, ..., h_{n_h});
for b in 1:B̃ do
    // sample
    Initialize Y ∈ ℝ^{(h_max + h̃)T};
    for k in 0:(h_max+h̃-1) do
        Sample j uniformly from (1, ..., N);
        Y[kT + 1 : kT + T] ← (x_{j1}, ..., x_{jT});
    // calculate p-value at any time for any lookback horizon
    for k in 0:(h̃-1) do
        for τ in 1:d:T do
            for h in h_1, ..., h_{n_h} do
                y ← Y[(h_max + k − h)T : (h_max + k)T + τ];
                p ← Individual_test(x_0 = x, x = y, s = s, B = B, BS = BS, α = 1).p;        (Algorithm 3)
                P[b] ← min(P[b], p);
Return P;
```

Algorithm 5: Sequential_degradation_test

Input: reference episodic signal $x_0 \in \mathbb{R}^{N \times T}$; test data stream x; test-statistic function s; inner-bootstrap repetitions $B \in \mathbb{N}$; tests frequency $d \in [1, T]$ and lookback horizons $h_1, ..., h_{n_h} \in \mathbb{N}$; family-wise significance parameters $\alpha_0 \in (0, 1), \tilde{h} \in \mathbb{N}$; outer-bootstrap repetitions $\tilde{B} \in \mathbb{N}$;

Output: time of H_0 rejection;

Algorithm:

```
Initialize bootstrap-storage BS;
P ← Sequential_bootstrap(x_0, s, B, BS, d, {h_i}, h̃, B̃);        (Algorithm 1)
α ← quantile_{α_0}(P);
if α = 1/(B+1) then
    // Can never reject H_0
    Warn("Either increase B or reduce significance requirements.");
    Return ERROR;
for k in (h_max, h_max+1, ...) do
    for τ in 1:d:T do
        for h in h_1, ..., h_{n_h} do
            y ← x[(k − h)T : kT + τ];
            r ← Individual_test(x = y, s = s, BS = BS, α = α).reject;        (Algorithm 3)
            if r=1 then
                // Reject H_0
                Return kT + τ;
```

H. Experiments Implementation Details

In the Pendulum (OpenAI) environment, where the goal is to keep a one-dimensional pendulum-pole pointing upwards, we define several alternative scenarios. *ccostx* scenario (with a parameter x) increases the cost of action ("control cost", which is quadratic in the activated force) to $x\%$ of its original value. Note that the control cost is typically the smaller among the components of the reward, which also include the angle of the pendulum and its speed. *noisex* scenario adds an additive

Table 2: Environments scenarios

Environment	Scenarios
Pendulum-v0	H_0 ccost$x$: action cost $\times$= x% noise$x$: additive noise = x% of max action len$x$: lenth $\times$= x% mass$x$: mass $\times$= x%
HalfCheetah-v3	H_0 ccost$x$: action cost $\times$= x% mass$x$: mass $\times$= x% gravity$x$: gravity $\times$= x%
Humanoid	H_0 ccost$x$: action cost $\times$= x% len$x$: leg size $\times$= x%

random normally-distributed noise to each action, whose standard deviation is $x\%$ of the range of valid actions. *lenx* and *massx* scenarios change the Pendulum length and mass respectively to $x\%$ of their original values. Note that while these scenarios are not necessarily harder to act in, the changes are still supposed to cause degradation since the agent is not re-trained for them.

In the HalfCheetah (MuJoCo) environment, where the goal is to train a two-dimensional cheetah to run as fast as possible, we also define several alternative scenarios. *ccostx* and *massx* are similar to the analog scenarios in Pendulum described above. The "control cost" in this case is also quadratic with the activated force, and is typically smaller than the other component of the reward – the speed of the Cheetah. *gravityx* scenario changes the gravity to $x\%$ of its original value.

Table 2 briefly summarizes the various scenarios, and Table 1 (in Section 6.1) summarizes the parameters of the tests setup per environment.

For every environment, before running the statistical tests according to Section 6.1, the recorded rewards are downsampled in time by factor d: every interval of samples $\{x_t\}_{t=d\cdot\tilde{t}+1}^{d\cdot\tilde{t}+d}$ is replaced by its mean as a single sample $\tilde{x}_{\tilde{t}} := \frac{1}{d}\sum_{t=d\cdot\tilde{t}+1}^{d\cdot\tilde{t}+d} x_t$. The downsampling reduces the dimension of the covariance matrix Σ_0 to $\frac{T}{d} \times \frac{T}{d}$, making it less noisy to estimate. In addition, it reduces the computational complexity of the experiments in this section. After the downsampling, the sequential tests apply an individual test in every single time-step, i.e., the testing-frequency of the sequential tests is $F = T/d$ per episode.

The statistical tests compared in Section 6.2 are mostly based on the threshold-tests described in Algorithm 3 (for individual tests) and Algorithm 5 (for sequential tests), with different test-statistics:

- **Mean**: simple mean of the rewards.
- **CUSUM**: the standard cumulative-sum (Page, 1954; NCSS) test, with every time-step normalized by its standard-deviation (estimated over all the reference episodes), and with reference value $k = 0.5$. As CUSUM is online by nature, it is used as is (beginning to run h episodes in advance for any lookback horizon h) instead of as part of Algorithm 5. The family-wise significance level of CUSUM is controlled using Algorithm 1. Note that through the normalization mentioned above, we let CUSUM take advantage of the episodic setup and the trusted reference data. Other normalization methods (time-invariant normalization and no normalization) did not improve CUSUM results in the experiments of Section 6.2.
- **Hotelling**: Hotelling test (Hotelling, 1931) is an optimal test for detection of mean-shift under unchanged covariance in multivariate normal variables. We generalize the implementation so that the test can be applied to non-integer number of multivariate variables (corresponding to non-integer number of episodes). Note that the alternative hypothesis of Hotelling test is very general, which results in reduced test power for the specific domain of interest in the current work – degradation – as demonstrated in Section 6.2.
- **UDWM**: the uniform-degradation weighted-mean from Definition D.6. The test based on this statistic is also named

UDT.

- **PDM**: 0.9-partial-degradation mean from Definition D.11, corresponding to degradation focused on 90% of the time-steps. The test based on this statistic is also named PDT. While this statistic may look quite similar to UDWM, it can produce different results in environments where UDWM concentrates most of the weights in few time-steps – if PDM "drops" these time-steps. This may make PDM more robust to degradation scenarios where the dominantly-weighted time-steps are not affected. Further discussion regarding p and the relation to the CVaR statistic is provided in Appendix D.
- **Mixed**: The mixed statistic essentially incorporates multiple statistics together – Mean and PDM in this case – and testing whether any of them has "extreme" values. It is defined as $s = min$(p-value(Mean), p-value(PDM)), i.e., we calculate both statistics and take the more significant p-value. Note that s itself is the statistic, and that its p-value is derived using Algorithm 2's bootstrap as in any of the other test statistics. We see below that the Mixed test enjoys most of the value of PDM, and still performs reasonably well wherever PDM is not robust enough (namely, when the mean reward decreases but the highly-weighted time-steps actually *increase*). The test based on this statistic is also named MDT.

I. Complementary Figures

Figures 5-14 introduce additional results from the experiments described in Section 6. Note that Section 6 refers directly to some of the results presented in this section.

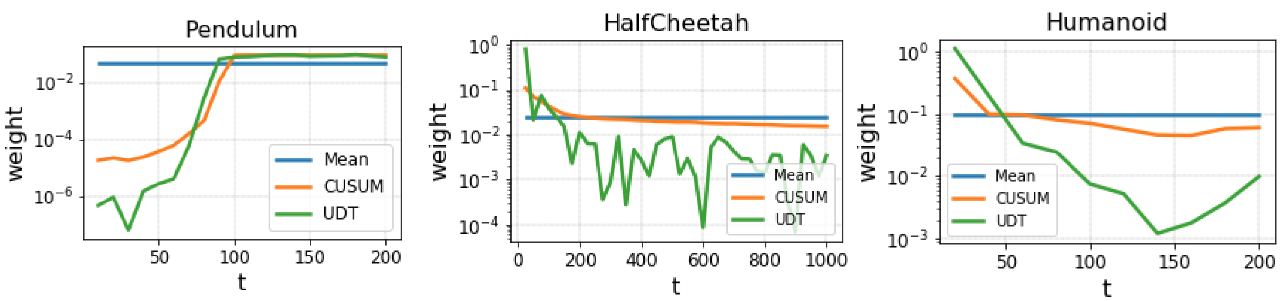


Figure 5: The weights assigned to the various time-steps by the various tests. Mind the logarithmic scale. Note that the weights of CUSUM are received from its normalization scheme, i.e., $w_t = 1/\text{std}(r_t)$.

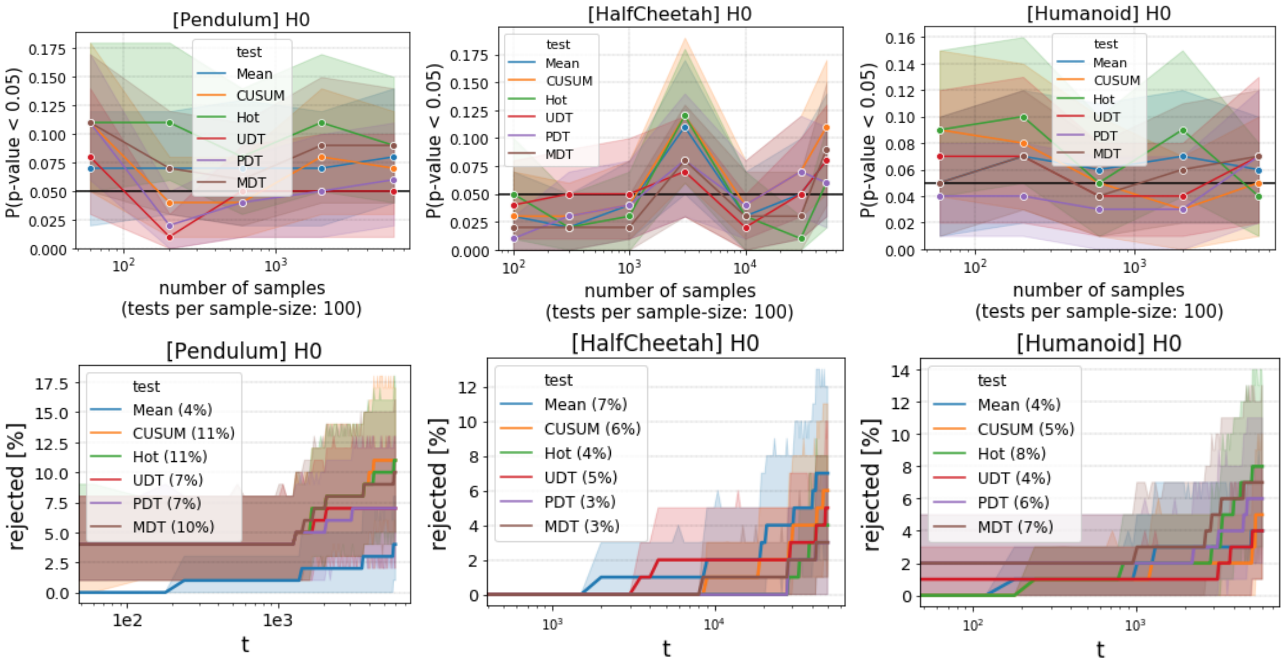


Figure 6: Percent of rejections of H_0 when H_0 is true, for various statistical tests, for both individual (top) and sequential (bottom) tests. Each point in each plot represents $M = 100$ tests, and the shaded area represents 95% confidence interval. The tests were tuned by Algorithm 2 (individual) and Algorithm 1 (sequential), using a reference dataset, to yield rejection rate of 5% under H_0.

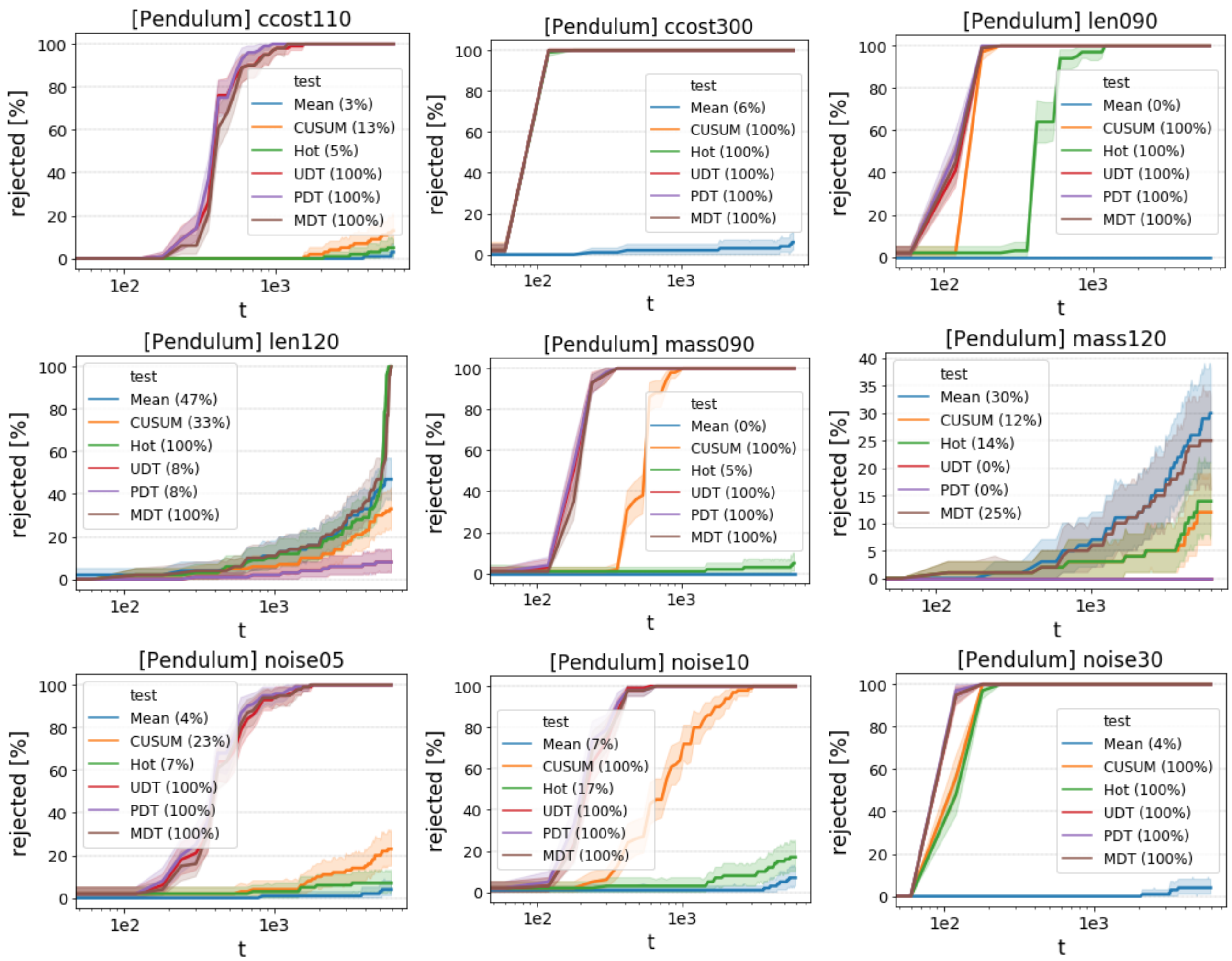


Figure 7: Sequential tests in different scenarios in Pendulum environment: cumulative percent of rejections vs. number of simulated time-steps. In the legend, the numbers in parenthesis are the final percents of rejection.

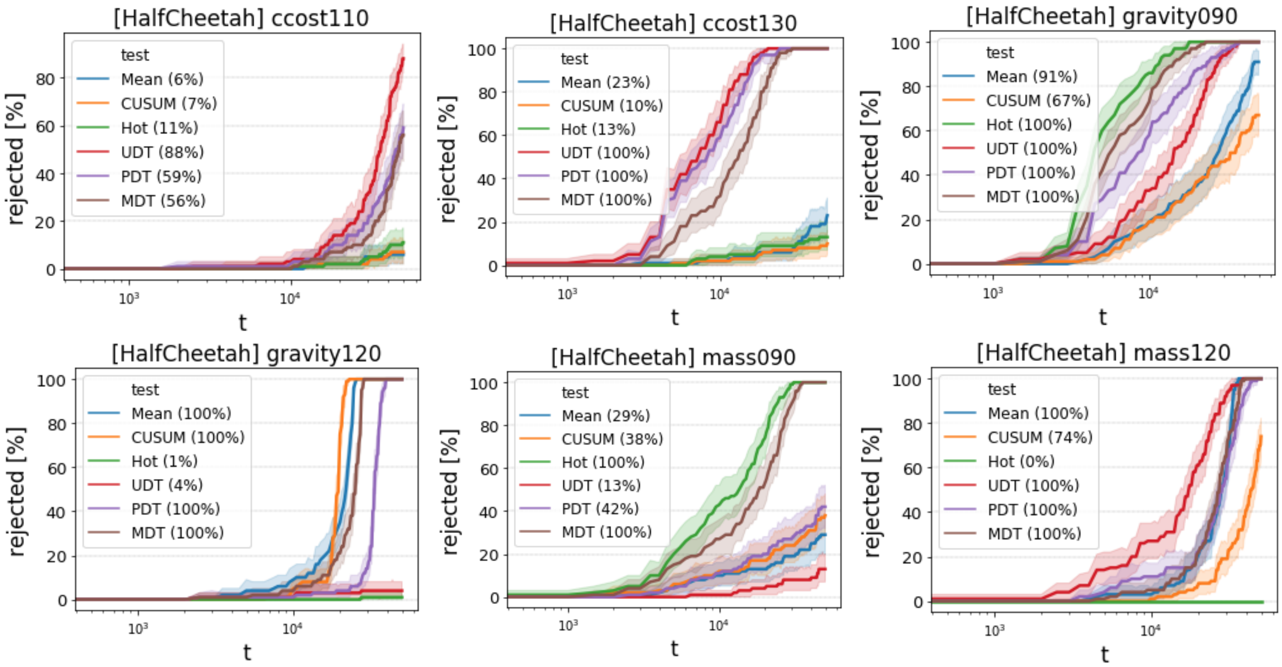


Figure 8: Sequential tests in different scenarios in HalfCheetah environment: cumulative percent of rejections vs. number of simulated time-steps. In the legend, the numbers in parenthesis are the final percents of rejection.

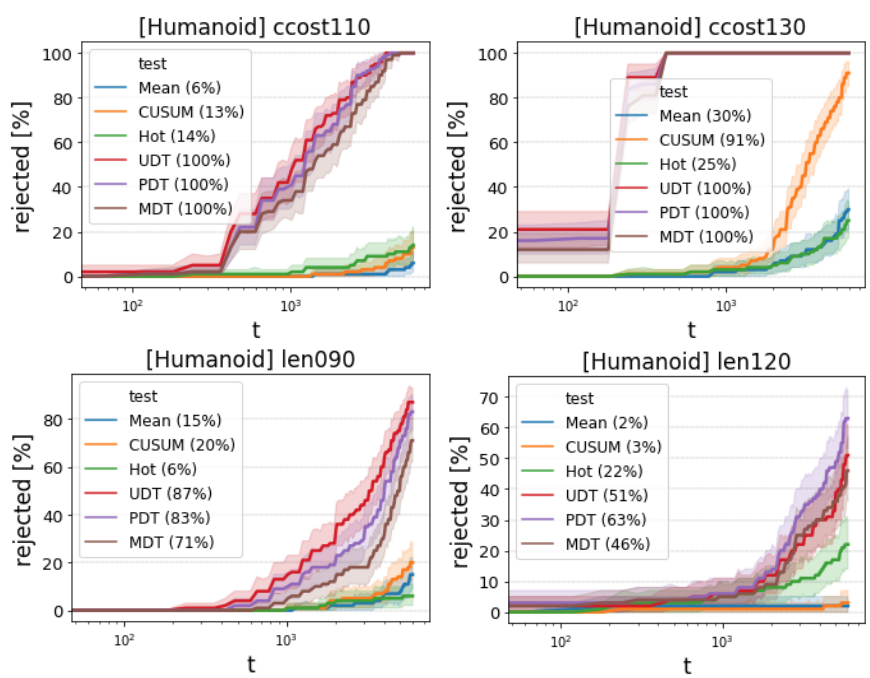
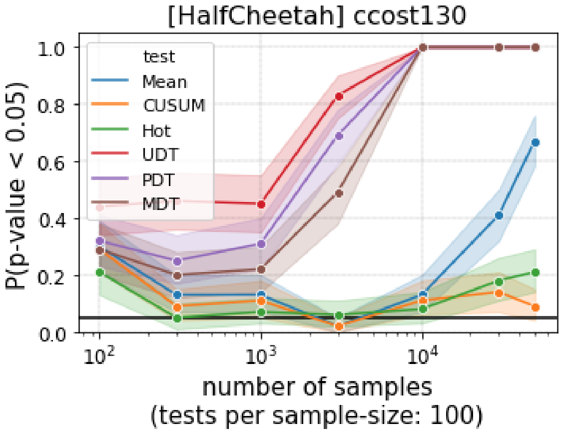
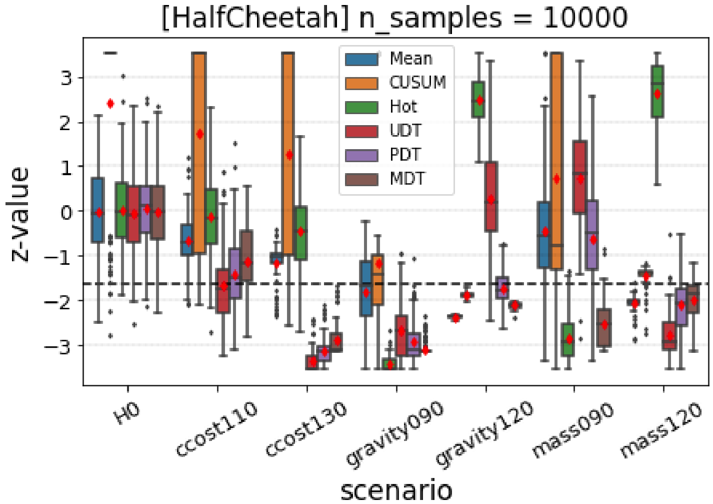


Figure 9: Sequential tests in different scenarios in Humanoid environment: cumulative percent of rejections vs. number of simulated time-steps. In the legend, the numbers in parenthesis are the final percents of rejection.



(a)



(b)

Figure 10: Individual (not sequential) tests in HalfCheetah environment: (a) percent of rejections (with significance $\alpha = 0.05$) vs. number of samples: recall that $T = 1000$ samples correspond to a single episode, and note that Mean, CUSUM and Hotelling perform better in the beginning of the first episode – before most of the noise comes in; (b) for each scenario and each test-statistic -– the distribution of the $M = 100$ z-values corresponding to simulated data blocks of 10 episodes each. The horizontal line represents the rejection threshold for significance $\alpha = 0.05$.

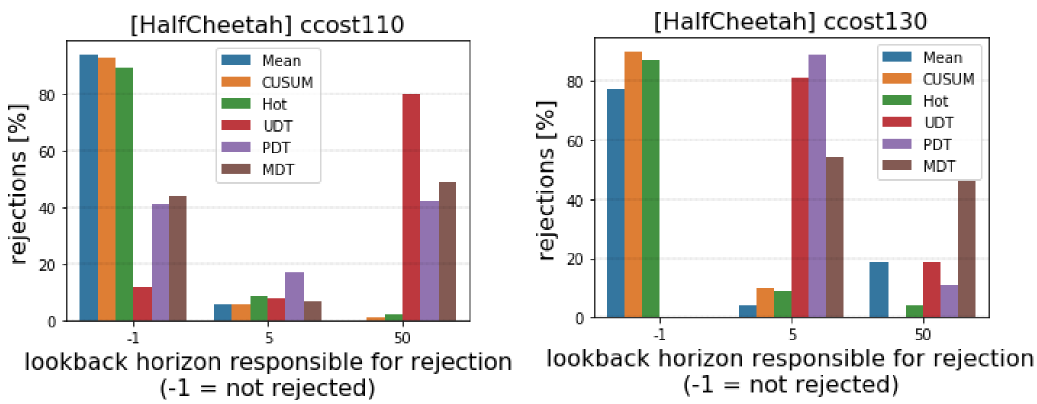


Figure 11: Lookback horizons for which H_0 was rejected in sequential tests: smaller degradation requires longer horizon (i.e., more data) for detection.

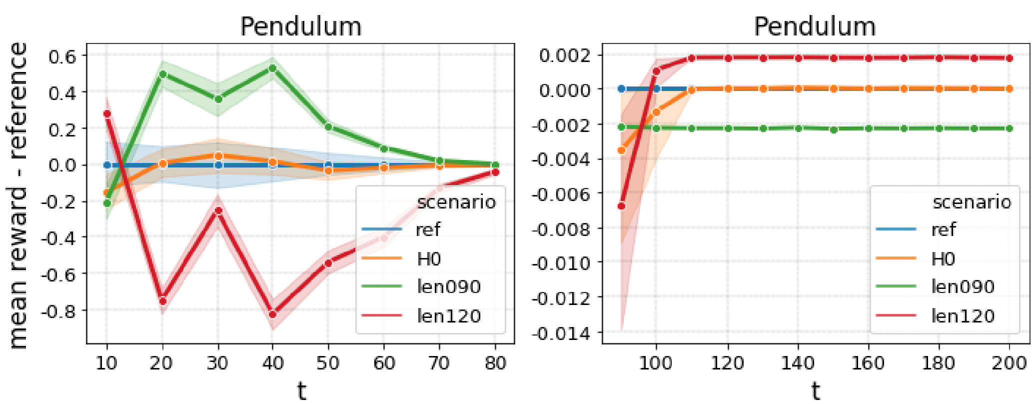
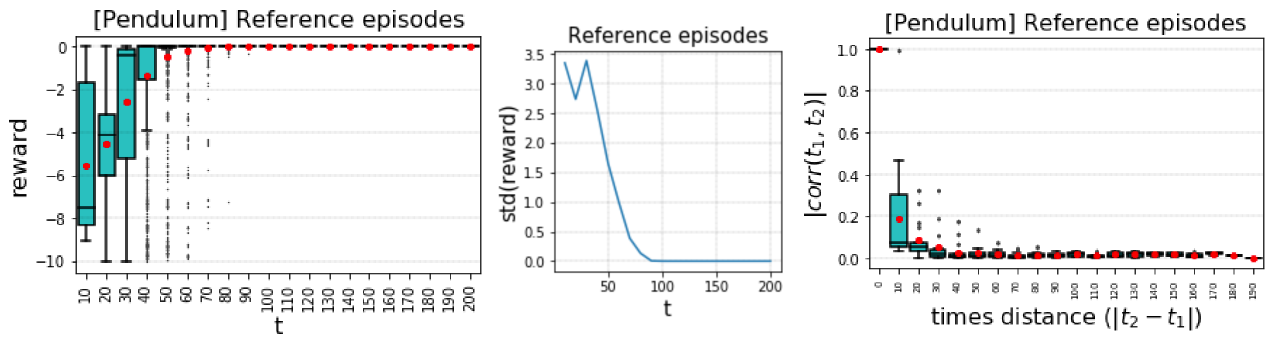
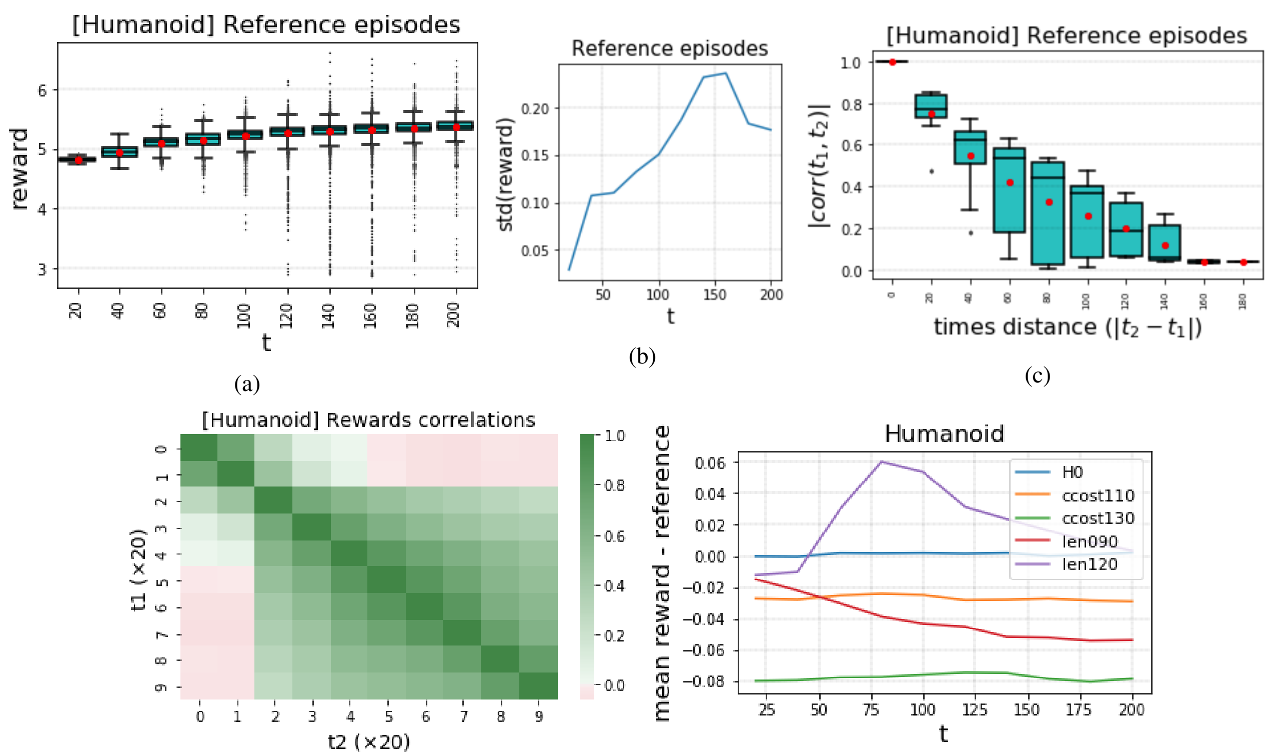


Figure 12: Rewards degradation in Pendulum following changes in pole length, over $N = 3000$ episodes per scenario. The figure is split due to the extreme scale difference between $t < 90$ and $t > 90$.



(a) (b) (c)

Figure 13: Parameters of an episodic signal of the rewards in Pendulum environment, estimated over $N = 3000$ episodes of $T = 200$ time-steps: (a) distribution of rewards per time-step; (b) standard deviations; (c) correlation(t_1, t_2) vs. $|t_2 - t_1|$. The estimations were done in resolution of 10 time-steps, i.e., every episode was split into 20 intervals of 10 consecutive rewards, and each sample is the average over an interval.



(a) (b) (c)

(d) (e)

Figure 14: Parameters of an episodic signal of the rewards in Humanoid environment, estimated over $N = 5000$ episodes of $T = 200$ time-steps: (a) distribution of rewards per time-step; (b) standard deviations; (c) correlation(t_1, t_2) vs. $|t_2 - t_1|$; (d) correlations map; (e) rewards degradation following changes in control costs and leg size. The estimations were done in resolution of 20 time-steps, i.e., every episode was split into 10 intervals of 20 consecutive rewards, and each sample is the average over an interval.

J. Sensitivity to Covariance Matrix Estimation

In most of the analysis in this work we assume that both the means $\boldsymbol{\mu_0}$ and the covariance Σ_0 of the episodic signal X are known. In practice, this can be achieved either through detailed domain knowledge, or by estimation from the recorded reference dataset of Setup 1, assuming it satisfies Eq. (1). The parameters estimation errors decrease as $\mathcal{O}(1/\sqrt{N})$ with the number N of reference episodes, and are distributed according to the Central Limit Theorem (for means) and Wishart distribution (K. V. Mardia & Bibby, 1979) (for covariance).

If N is suspected to be too small for accurate estimation, it is possible to deal with the estimation errors of the model parameters through regularization. One possible regularization is assuming absence of correlations between distant time-steps ($\exists \delta \in \mathbb{N}, \forall |t_2 - t_1| > \delta : (\Sigma_0)_{t_1 t_2} = 0$). Another is to essentially reduce T through grouping of sequences of time-steps together (as we do in Section 6, for example).

To test the practical consequences of inaccurate parameters estimation, we repeated some of the offline (individual) tests of Section 6 for HalfCheetah – with different sizes of reference datasets. The reference datasets vary between $N = 100$ and $N = 10000$ episodes (where $N = 10000$ corresponds to Section 6). As in Section 6, we downsample each episode from $T = 1000$ to $F = T/d = 40$ time-steps.

Figure 15 shows the results of the sensitivity tests. Even with as little as $N = 100$ reference episodes, the largest weights are successfully assigned to the first time-steps (mind the logarithmic scale in both axes), although certain later weights are still noisy. $N = 300$ is sufficient to yield a consistent statistic distribution under H_0, i.e., to reliably tune the false alarm rate. All sizes of reference datasets yield similar test power in the tested scenarios *ccost130* and *gravity090*. $N = 3000$ is hardly distinguishable from $N = 10000$ by any mean.

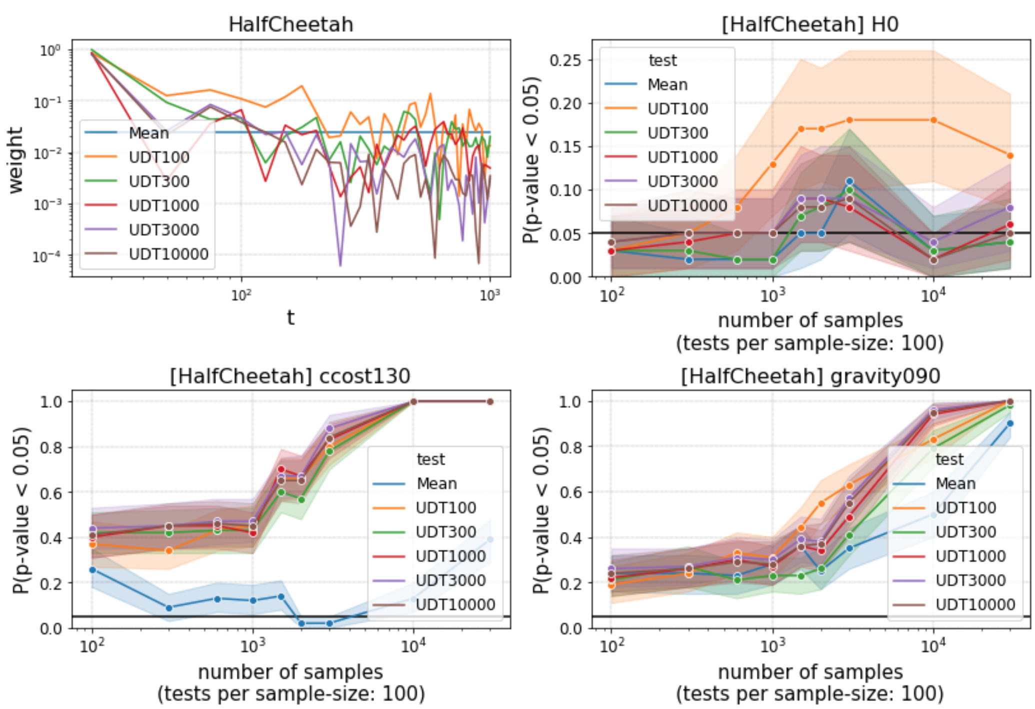


Figure 15: The weights of Uniform Degradation Tests (UDT), based on estimation of parameters from reference datasets of various sizes (top left); and percents of degradation detections in individual tests in different scenarios (with significance $1 - \alpha = 0.95$).