# Reinforcement Learning Design for Quickest Change Detection

Austin Cooper and Sean Meyn*

*Abstract*— **The field of quickest change detection (QCD) concerns design and analysis of algorithms to estimate in real time the time at which an important event takes place, and identify properties of the post-change behavior.**

**It is shown in this paper that approaches based on reinforcement learning (RL) can be adapted based on any "surrogate information state" that is adapted to the observations. Hence we are left to choose both the surrogate information state process and the algorithm. For the former, it is argued that there are many choices available, based on a rich theory of asymptotic statistics for QCD. Two approaches to RL design are considered:**

**(i) Stochastic gradient descent based on an actor-critic formulation. Theory is largely complete for this approach: the algorithm is unbiased, and will converge to a local minimum. However, it is shown that variance of stochastic gradients can be very large, necessitating the need for commensurately long run times.**

**(ii) Q-learning algorithms based on a version of the projected Bellman equation. It is shown that the algorithm is stable, in the sense of bounded sample paths, and that a solution to the projected Bellman equation exists under mild conditions.**

**Numerical experiments illustrate these findings, and provide a roadmap for algorithm design in more general settings.**

## I. INTRODUCTION

The goal of the research surveyed in this paper is to create algorithms for *quickest change detection* (QCD), for applications in which statistics are only partially known, particularly after the change has occurred. While the authors were initially motivated by applications in power systems, the setting here is entirely general. Examples of events that we wish to detect include human or robotic intruders, computer attack, faults in a power system, and onset of heart attack for a patient [12], [13].

The standard QCD model includes a sequence of observations $\boldsymbol{Y} := \{Y_k : k \geq 0\}$, assumed here to evolve as a real-valued stochastic process. The statistics of these observations change at a time denoted $\tau_{\mathsf{a}} \geq 0$. The goal is to construct an estimate of the change time, denoted $\tau_{\mathsf{s}}$, that is adapted to the observations. The estimate must balance two costs: 1. *Delay*, which is expressed $(\tau_{\mathsf{s}} - \tau_{\mathsf{a}})_+ := \max(0, \tau_{\mathsf{s}} - \tau_{\mathsf{a}})$, and 2. *false alarm*, meaning that $\tau_{\mathsf{s}} - \tau_{\mathsf{a}} < 0$.

There are two general models that lead to practical solutions: Bayesian and minimax approaches. Typical measures of performance for the former approach are based on *mean detection delay* MDD and probability of false alarm $p_{\mathsf{FA}}$:

$$\mathrm{MDD} = \mathsf{E}[(\tau_{\mathsf{s}} - \tau_{\mathsf{a}})_+] \quad \text{and} \quad p_{\mathsf{FA}} = \mathsf{P}\{\tau_{\mathsf{s}} < \tau_{\mathsf{a}}\}. \quad (1)$$

The focus of this paper is on the Bayesian approach, based on a partially observed Markov Decision Process (POMDP). See Section II-A for canonical examples.

Successful approaches to algorithm design are typically based on the construction of a real-valued stochastic process $\{\mathcal{X}_n\}$ that plays a role similar to the celebrated information state of POMDP theory, and a threshold policy is adopted: for a pre-assigned threshold $\mathsf{H} > 0$, the stopping rule is

$$\tau_{\mathsf{s}} = \min\{n \geq 0 : \mathcal{X}_n \geq \mathsf{H}\}. \quad (2)$$

Two famous examples are defined recursively: with $\mathcal{X}_0 = 0$,

1. Shiryaev–Roberts: $\quad \mathcal{X}_{n+1} = \exp(L_{n+1})[\mathcal{X}_n + 1]$ (3a)
2. CUSUM: $\quad\quad\quad\quad \mathcal{X}_{n+1} = \max\{0, \mathcal{X}_n + L_{n+1}\}$ (3b)

in which $L_n = L(Y_n)$ is a log likelihood ratio for the conditional i.i.d. settings in which these models are typically posed (see Section II). In particular, the CUSUM statistic evolves as a reflected random walk (RRW) with negative drift for $0 \leq n < \tau_{\mathsf{a}}$.

Analysis of the threshold policy (2) is typically posed in an asymptotic setting, considering a sequence of models with threshold $\mathsf{H}$ tending to infinity. Approximate optimality results for either statistic may be found in [16], [17]. See [18], [24] for further history.

**Contributions** This paper develops theory for QCD in a Bayesian setting, and demonstrates how the solution structure lends itself to RL design. One theme is the application of observation-driven statistics such as (3b) to form a "surrogate" information state for policy synthesis.

● Performance of the CUSUM test is approximated in the asymptotic setting in which there is a strong penalty for false alarm. Well-known analysis for the ideal case is extended to settings for which $L_n$ in (3b) is mismatched with the observations (i.e., $L_n$ is not a log likelihood ratio). This general theory motivates the control architectures proposed for RL design. In particular:

▷ An actor-critic approach is introduced and shown to be consistent under mild conditions (see Prop. 3.2).

▷ A Q-learning algorithm is introduced and shown to be stable provided the input used for training is *sufficiently optimistic* [15].

● The theory is illustrated with many experiments, comparing resulting policies with common heuristics as well as the true optimal. Among the findings are

▷ Stability for the scalar gain algorithm requires extremely high level of optimism, resulting in poor numerical performance. A version of Zap Q-learning is far more reliable.

▷ The resulting policies performed well using a basis obtained via binning, and a linear function class inspired by results obtained via binning.

**Literature** See [12], [13] for excellent recent surveys on QCD theory. Much of this theory is cast in a minimax rather than Bayesian setting. Numerical techniques to solve the QCD problem in the Bayesian setting may be found in [27].

The analysis in Section II-B is cast as the conditionally i.i.d. model of [19], though we relax the classical assumption that the change time has a geometric distribution.

The vast majority of theory requires statistical independence of $X^0$, $X^1$ and $\tau_a$. This is the case in Shiryaev's conditional i.i.d. model, for which the stochastic processes $X^0$, $X^1$ are also assumed i.i.d.;

Extension to a conditionally Markov model or hidden Markov model is possible by adapting techniques from the recent work [26], [25]; while cast in an adversarial setting, many approximations remain valuable in the Bayesian setting adopted here.

Stability theory of Q-learning for optimal stopping was resolved in [21]; conditions for consistency are similar to those for the simpler TD-learning algorithm. However, the specific algorithm considered required that the cost function be fully observed. This is why Q-learning is re-considered in the present article.

In this prior work it is recognized that the state is inherently partially observed. In [21] along with many papers in the RL literature, a truncated history of observations is adopted as a surrogate information state, $\mathcal{X}_k = (Y_{k-B+1}; \cdots; Y_k)$, with $B \geq 1$. An innovations process obtained from the Kalman filter is used to define $\{\mathcal{X}_k\}$ in applications to power systems [10]. General theory surrounding the approximation of the information state may be found in [20] (along with substantial history).

There is a long history of application of techniques from RL to approximate the solution to the optimal stopping problem. The first stability analysis of Q-learning with linear function approximation appeared in [21], which inspired significant research such as [11], [2]. The algorithms conceived in this prior work are not applicable in the applications considered in this paper because the cost (or rewards) is assumed to be fully observed. The RL algorithms introduced in this paper are more complex, and have a weaker supporting stability theory, precisely because this assumption is violated.

**Organization** Section II provides background on the standard Bayesian QCD problem as well as an alternate cost criterion for which approximations are formulated. Section III includes formulations of two RL approaches to optimal stopping. The paper then turns to design and experimental findings of Q-learning applied to our Bayesian QCD problem in Section IV. Section V provides concluding thoughts and directions for future work.

## II. BAYESIAN QCD

This section contains background on approaches to modeling and algorithm design for QCD. We begin with a canonical Bayesian model, cast as a POMDP.

### A. POMDP model

In this model both the change time and the observations are deterministic functions of a time-homogeneous Markov chain $\Phi$, evolving on a state space $X$. It is assumed that $Y_k = h(\Phi_k)$, $k \geq 0$, for a function $h \colon X \to Y$ (measurable in an appropriate sense). Assume moreover that there is a decomposition $X = X_0 \cup X_1$, for which $X_1$ is *absorbing*: $\Phi_k \in X_1$ for all $k \geq 0$ if $\Phi_0 \in X_1$. The change time is defined by $\tau_a = \min\{k \geq 0 : \Phi_k \in X_1\}$.

We arrive at a POMDP with input $U_k \in U = \{0, 1\}$, and $\tau_s$ defined as the first value of $k$ such that $U_k = 1$. The control problems of interest are optimal stopping problems: For any cost functions $c_\circ, c_\bullet \colon X \to \mathbb{R}$, we wish to minimize over all inputs adapted to the observations,

$$J(\Phi_0, U_0^\infty) = \mathsf{E}\Big[ \sum_{k=0}^{\tau_s - 1} c_\circ(\Phi_k) + c_\bullet(\Phi_{\tau_s})\Big] \qquad (4)$$

Consistent with the standard QCD framework is $c_\circ(z) = \mathbf{1}\{z \in X_1\}$ and $c_\bullet(z) = \kappa\mathbf{1}\{z \in X_0\}$ with $\kappa > 0$, so that $J(\Phi_0, U_0^\infty) = \mathsf{MDD} + \kappa p_{\mathsf{FA}}$ (recall (1)).

The structure of an optimal solution can be expressed as state feedback with suitable choice of state process. We use the term *information state*, denoted $\{\mathcal{X}_k : k \geq 0\}$. This is defined as a sufficient statistic for optimal control, in the sense that an optimal solution is expressed as "information state feedback", $U_k^* = \phi^*(\mathcal{X}_k)$. The canonical example is $\{\mathcal{X}_k\} = \{\Pi_k\}$, the sequence of conditional distributions (often called the belief state) [7], [9]. This structure leads to a practical solution when $X$ is finite, with $K$ elements, so that $\Pi_k$ evolves on the $K$-dimensional simplex $\mathcal{S}^K$, and $\phi^* \colon \mathcal{S}^K \to U$ is measurable.

**Shiryaev's model** The POMDP model is a generalization of Shiryaev's conditional i.i.d. model, in which observations are expressed

$$Y_k = X_k^0 \mathbf{1}_{k < \tau_a} + X_k^1 \mathbf{1}_{k \geq \tau_a}, \qquad k \geq 0, \qquad (5)$$

with $X^0$ and $X^1$ i.i.d. and mutually independent stochastic processes; the change time $\tau_a$ is independent of $X^0, X^1$, and has a geometric distribution. Under these strong assumptions, the real-valued process $\{p_k = \mathsf{P}\{\tau_a \leq k \mid Y_0^k\} : k \geq 0\}$ serves as an information state, and an optimal test is of the form $U_k^* = \mathbf{1}\{p_k \geq \mathsf{H}\}$ for some threshold $\mathsf{H} > 0$ (see [19] and the tutorial [23]).

The observation model (5) is valuable in analysis of common heuristics. Suppose that the marginal distributions of $\{X_k^0, X_k^1\}$ have densities on $\mathbb{R}$, denoted $f_0$, $f_1$, and denote $L(y) = \log(f_1(y)/f_0(y))$—the log likelihood ratio (LLR). Crucial for analysis of either of the algorithms (3) is that $L$ has positive mean under $f_1$ and negative mean under $f_0$.

It is known that either of the algorithms (3) is approximately optimal for Shiryaev's model, for large $\kappa$ and large mean change time $\mathsf{E}[\tau_a]$ [24].

**Alternative to the standard cost criterion** The standard cost criterion is $\mathsf{MDD} + \kappa p_{\mathsf{FA}}$ is sensible in Shiryaev's model in which the change time is independent of $\{X_k^0, X_k^1 : k \geq$

0}. In the general POMDP model there may be evidence that a change is imminent; in such cases, a (common sense) good decision rule might make an early declaration of change. These decision rules might be far from optimal under the usual cost criterion since it is insensitive to the value of *eagerness*, defined as $(\tau_s - \tau_a)_- := \max(0, -[\tau_s - \tau_a])$. In this paper we consider the *mean detection eagerness* $\mathsf{MDE} = \mathsf{E}[(\tau_s - \tau_a)_-]$ in the cost criterion $\mathsf{MDD} + \kappa\mathsf{MDE}$, leading to what we believe is a more reasonable objective,

$$J(\Phi_0, U_0^\infty) = \mathsf{E}\big[(\tau_s - \tau_a)_+ + \kappa(\tau_s - \tau_a)_-\big] \quad (6)$$

This may be placed in the POMDP standard form (4), with

$$c_\circ(z) = \mathbf{1}\{z \in \mathsf{X}_1\}, \qquad c_\bullet(z) = \kappa\mathsf{E}[\tau_a \mid \Phi_0 = z] \quad (7)$$

### B. Asymptotic statistics

The remainder of this section concerns the CUSUM test. Analysis is restricted to the following conditional i.i.d. setting: The stochastic processes $\{X_k^0\}$, $\{X_k^1\}$ are each i.i.d., and independent of the change time $\tau_a$. It is assumed that the marginal densities $f_0$ and $f_1$ exist, and that the LLR $L = \log(f_1/f_0)$ exists and is integrable with respect to either $f_1$ or $f_0$

Our interest is approximating the performance of the CUSUM test, and also approximating the optimal threshold for a given value of $\kappa$. The analysis allows for two significant relaxations:

**1.** We consider $L_n = F(Y_n)$ for a Borel measurable function $F: \mathsf{Y} \to \mathbb{R}$, not necessarily the LLR. Letting $m_i = \int F(y) f_i(y) dy$ for $i = 0, 1$, it is assumed that $m_0 < 0$ and $m_1 > 0$. Hence the RRW (3b) is a positive recurrent Markov chain if $\tau_a = \infty$.

Two log moment generating functions are denoted $\Lambda_i(\upsilon) = \log\mathsf{E}[\exp(\upsilon F(X_k^i))]$ for $\upsilon \in \mathbb{R}$ and $i = 0, 1$.

**2.** The strong distributional assumption on the change time is replaced by a regularity condition:

**Regular geometric tail:** for some $\varrho_a < \infty$,

$$\lim_{n\to\infty} \frac{1}{n} \log\mathsf{P}\{\tau_a \geq n\} = -\varrho_a \quad (8)$$

The regularity assumption obviously holds in Shiryaev's model, in which $\tau_a$ has a geometric distribution. We obtain $\varrho_a > 0$ in the POMDP model under mild assumptions. Full proofs for results in this section may be found in [4], [5], along with extensions beyond the conditionally i.i.d. model.

*Lemma 2.1:* Consider the POMDP model with $\mathsf{X}$ finite, $\mathsf{P}\{\tau_a = \infty\} = 0$, yet $\mathsf{P}\{\tau_a > N\} > 0$ for each $N > 0$ and $\Phi_0 \in \mathsf{X}_1$. Then (8) holds for some $\varrho_a > 0$.

The cost of delay is easily approximated for this model: After a change has occurred, the most likely path is linear with slope $m_1 > 0$. For a threshold $\mathsf{H} \gg 1$, the delay $(\tau_s - \tau_a)_+$ is overwhelmingly likely to be close to $\mathsf{H}/m_1$.

Approximation of the mean of $(\tau_s - \tau_a)_-$ is based on well-established large deviations theory for RRWs. The main results of this theory require that the log moment generating functions $\Lambda_i(\upsilon) := \log\int \exp(\upsilon x) f_i(x)\, dx$, $i = 0, 1$, be finite over a suitable range of $\upsilon \in \mathbb{R}$.

Denote for any threshold test,

$$\mathsf{H}^*(\kappa) = \arg\min_{\mathsf{H}\geq 0}\{\kappa\mathsf{MDE}(\mathsf{H}) + \mathsf{MDD}(\mathsf{H})\} \quad (9a)$$

$$J^*(\kappa) = \min_{\mathsf{H}\geq 0}\{\kappa\mathsf{MDE}(\mathsf{H}) + \mathsf{MDD}(\mathsf{H})\} \quad (9b)$$

We write CUSUM* to denote the CUSUM algorithm using the optimal threshold $\mathsf{H}^*(\kappa)$. In Prop. 2.2 we justify these approximations,

$$\bar{\mathsf{H}}_\infty^*(\kappa) = \frac{1}{\upsilon_+} \log(\kappa), \quad \bar{J}_\infty^*(\kappa) = \frac{1}{m_1}\frac{1}{\upsilon_+} \log(\kappa) \quad (10)$$

where the constant $\upsilon_+$ is defined in the proposition.

*Proposition 2.2:* Suppose the following conditions hold: 1) the limit (8) holds with $\varrho_a > 0$, 2) $\Lambda_0$ has two distinct roots $\{0, \upsilon_0\}$, a unique solution $\upsilon_+ > \upsilon_0$ to $\Lambda_0(\upsilon_+) = \varrho_a$, and $\Lambda_0$ is finite-valued in a neighborhood of $[0, \upsilon_+]$, 3) $\Lambda_1$ is finite-valued in a neighborhood of the origin. Then, for the CUSUM* test,

$$\mathsf{H}^*(\kappa) = \bar{\mathsf{H}}_\infty^*(\kappa) + o(\log(\kappa)) \quad (11a)$$

$$J^*(\kappa) = \bar{J}_\infty^*(\kappa) + o(\log(\kappa)) \quad (11b)$$
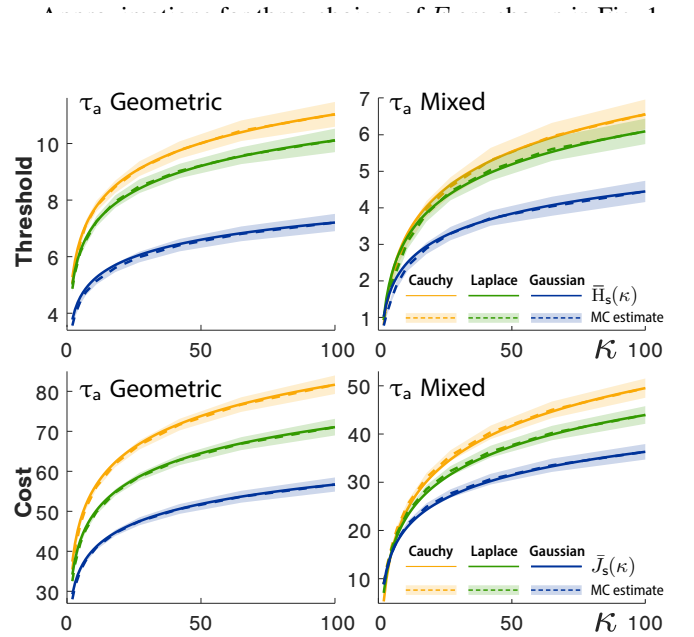
∎



Fig. 1. Approximations of the optimal threshold and cost.

### III. REINFORCEMENT LEARNING AND QCD

In the few examples we have considered we have found that the approximations in Prop. 2.2 are highly accurate. In non-ideal settings the proposition is valuable in the construction of RL algorithms. We provide algorithms, and full justification in some cases. Algorithm design and analysis is set in the POMDP (Bayesian) setting, with cost criterion (6).

Assumed given is a surrogate belief state: a stochastic process $\{\mathcal{X}_k : k \geq 0\}$, evolving on a closed subset of Euclidean space $\mathsf{S}$, and which is adapted to the observations $\mathcal{Y}_k = \sigma\{Y_0, \ldots, Y_k\}$. We do not require that $\mathcal{X}_k$ is in any
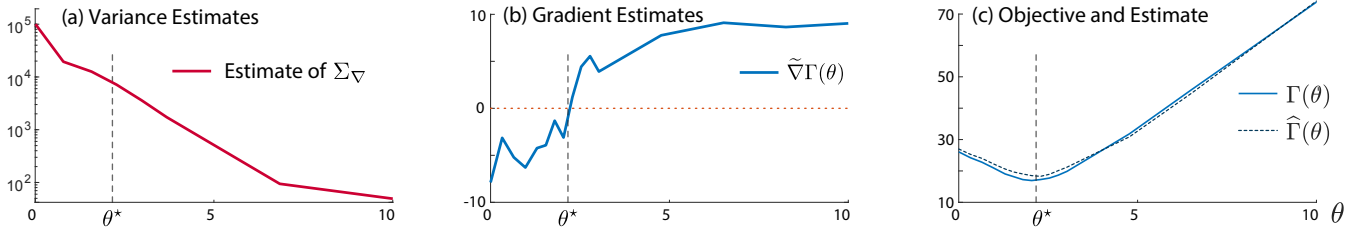
Fig. 2. Statistics of the gradient estimate as a function of $\theta$, based on the Actor Critic method: (a) Empirical variance of the gradient estimates. (b) Gradient estimates using $N = 10^4$ episodes. (c) Objective and its approximation obtained from integrating the gradient estimate.

sense an approximation of an information state. In particular, the numerical experiments largely focus on the CUSUM statistic (3b).

We first consider a version of the actor-critic method, followed by approaches to Q-learning.

**Actor-critic method** Assumed given is a collection of randomized stationary policies $\{\widetilde{\phi}^\theta : \theta \in \mathbb{R}^d\}$. For each $\theta$ the statistics of the decision rule are defined via

$$\mathsf{P}\{U_k = u \mid Y_0^k; \mathcal{X}_k = x\} = \widetilde{\phi}^\theta(u \mid x), \quad u \in \mathsf{U}, \; x \in \mathsf{S}.$$

We fix an initial distribution $\nu$ for the Markov chain $\Psi$, and denote $\mu_\theta(z, u) = \nu(z)\widetilde{\phi}^\theta(u \mid x)$ for $z = (x; \varsigma) \in \mathsf{X} \times \mathsf{S}$ and $u \in \mathsf{U}$. Our goal is to minimize

$$\Gamma(\theta) = \mathsf{E}_{\mu_\theta}^\theta\Big[\sum_{k=0}^{\tau_{\mathsf{S}}} c(\Phi_k, U_k)\Big] \tag{12}$$

The subscript indicates that $(\Phi_0, \mathcal{X}_0, U_0) \sim \mu_\theta$, and the superscript "$\theta$" indicates that the policy $\widetilde{\phi}^\theta$ determines the input.

We consider stochastic gradient descent (SGD),

$$\theta_{n+1} = \theta_n - \alpha_{n+1} G_n \breve{\nabla}_\Gamma(n), \tag{13}$$

in which the stepsize $\alpha_{n+1}$ and the matrix gain $G_n$ are design choices.

In the actor-critic algorithm the stochastic gradient $\breve{\nabla}_\Gamma(n)$ is represented in terms of the *score function*,

$$\Lambda^\theta(x, u) = \nabla_\theta \log[\widetilde{\phi}^\theta(u \mid x)] \tag{14}$$

defined to be zero for any values for which $\widetilde{\phi}^\theta(u \mid x) = 0$.

The following result follows from a long history surveyed in the Notes section of [14, Ch. 10]:

*Proposition 3.1:* Suppose that $\Gamma$ and $\nabla\Gamma$ are continuous. Then $\nabla\Gamma(\theta) = \mathsf{E}_{\mu_\theta}^\theta[\breve{\nabla}_\Gamma]$, for either of the two options:

$$\breve{\nabla}_\Gamma^\theta = \sum_{k=0}^{\tau_{\mathsf{S}}} c(\Phi_k, U_k) S_k^\theta \tag{15a}$$

$$or \quad \breve{\nabla}_\Gamma^\theta = \sum_{k=0}^{\tau_{\mathsf{S}}} Q_\theta(\Psi_k, U_k)\Lambda_k^\theta \tag{15b}$$

in which $\Lambda_k^\theta := \Lambda^\theta(\Phi_k, U_k)$, $S_k^\theta = \Lambda_0^\theta + \cdots + \Lambda_k^\theta$, and

$$Q_\theta(z, u) = \mathsf{E}^\theta\Big[\sum_{k=0}^{\tau_{\mathsf{S}}} c(\Phi_k, U_k) \mid \Psi_0 = z, U_0 = u\Big] \tag{15c}$$

With either representation for the stochastic gradient (15a) or (15b) we obtain an asymptotically unbiased SGD algorithm using $\breve{\nabla}_\Gamma(n) = \breve{\nabla}_\Gamma^{\theta_n}$. Each of those described here are episodic: data is collected over the period $0 \leq k \leq \tau_{\mathsf{S}}(n)$ with $\theta_n$ fixed, and the input defined using $\widetilde{\phi}^{\theta_n}$.

The *natural gradient descent* algorithm updates the matrix gain via $G_n = \widehat{R}_n^{-1}$, where $\widehat{R}_0 > 0$ with updates obtained recursively,

$$\widehat{R}_n = \widehat{R}_{n-1} + \beta_n[-\widehat{R}_{n-1} + R_n], \quad n \geq 1,$$

$$R_n = \sum_{k=0}^{\tau_{\mathsf{S}}(n)} \Lambda_k^{\theta_n}[\Lambda_k^{\theta_n}]^\mathsf{T}, \qquad \Lambda_k^{\theta_n} = \Lambda^{\theta_n}(\mathcal{X}_k, U_k) \tag{16}$$

with $\beta_n \gg \alpha_n$ (see [14, Ch. 10]).

The two representations for the gradients prompt two choices for the stochastic gradient. We focus here on the first,

$$\breve{\nabla}_\Gamma(n) = \sum_{k=0}^{\tau_{\mathsf{S}}(n)} c(\Phi_k, U_k) S_k^{\theta_n}$$

leaving out the extension of the standard algorithm based on TD(1) learning to estimate the Q-function [14, Ch. 10].

The Polyak-Ruppert (PR) estimates are defined by

$$\theta_n^{\mathsf{PR}} = \frac{1}{n} \sum_{i=1}^n \theta_i, \qquad n \geq 1. \tag{17}$$

Its *asymptotic covariance* is defined as

$$\Sigma_\Theta^{\mathsf{PR}} = \lim_{n \to \infty} n\, \mathsf{E}[\tilde{\theta}_n^{\mathsf{PR}}\{\tilde{\theta}_n^{\mathsf{PR}}\}^\mathsf{T}] \tag{18}$$

When this exists and is finite, then the estimates achieve the optimal mean-square convergence rate of $O(1/n)$.

The following is a consequence of recent stochastic approximation theory in [1].

*Proposition 3.2:* Suppose that the assumptions of Prop. 3.1 hold, and in addition (i) $\Gamma$ is coercive with unique minimum $\theta^*$ and $\nabla\Gamma$ is globally Lipschitz continuous. (ii) $A^* := \nabla^2\Gamma(\theta^*)$ is Hurwitz, and the steady-state covariance $R^* = \mathsf{Cov}(\Lambda^{\theta^*})$ is full rank. (iii) The stepsize sequence is $\alpha_n = \alpha_0 n^{-\rho}$ with $1/2 < \rho < 1$ and $\alpha_0 > 0$.

Then, the SGD algorithm (13) is convergent almost surely and in mean square. The PR-estimates are also convergent in both senses.

The Central Limit Theorem (CLT) holds, as well as the limit (18), in which the asymptotic covariance is $\Sigma_\Theta^{\mathsf{PR}} = [(R^*)^{-1}A^*]^\mathsf{T}\Sigma_{\breve{\nabla}}^*(R^*)^{-1}A^*$ with $\Sigma_{\breve{\nabla}}^*$ is the steady-state covariance of (15b) using the policy $\widetilde{\phi}^{\theta^*}$. ∎
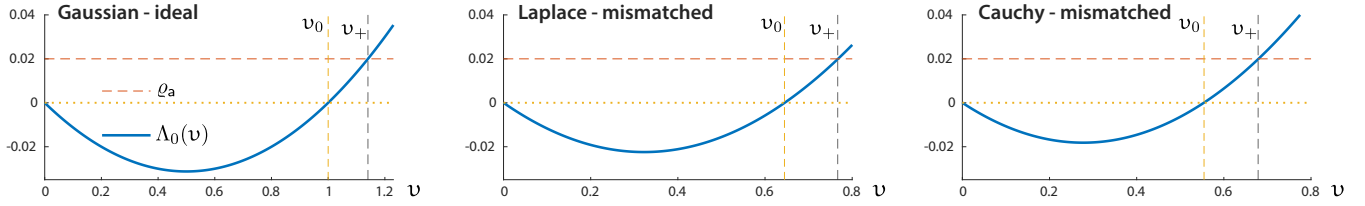
Fig. 3. $\Lambda_0(\upsilon)$ for ideal Gaussian alongside Laplace and Cauchy mismatched detectors

**Example** Consider the one-dimensional family of policies in which $\theta$ approximates a threshold rule: for a fixed large constant $\xi > 0$, define $\widetilde{\phi}^\theta(u \mid w) = [1 + \exp(\xi[w - \theta])]^{-1}\exp(\xi u[w - \theta])$, so that the score function is

$$\Lambda^\theta(u \mid w) = -\xi u + \xi\widetilde{\phi}^\theta(1 \mid w)$$

In this scalar example we can adapt the natural gradient actor critic method to estimate $\nabla\Gamma(\theta)$ for any fixed $\theta$.

Fig. 2 shows results from a typical experiment using $\xi = 20$. Details on the simulation environment designed to obtain these approximations are postponed to the Appendix.

Rather than demonstrate results from an application of SGD, the first two plots show estimates of the mean and variance of the random variable in the expectation (15b) for a range of values of $\theta$; the precise means are $\Sigma_\nabla(\theta)$ and $\nabla\Gamma(\theta)$.

Fig. 2 (b) shows gradient estimates, and Fig. 2 (c) compares estimates of the objective function $\Gamma(\theta)$ obtained via standard Monte-Carlo, and the estimate obtained from integrating the gradient estimates in (b): $\widehat{\Gamma}(\theta) := \kappa + \int_0^\theta \widehat{\nabla\Gamma}(r)\, dr$. The results indicates good news: in spite of the enormous variance shown in (a), especially large for smaller values of $\theta$, the zero of the gradient estimate $\widehat{\nabla\Gamma}(\theta)$ is very close to the optimal threshold value for CUSUM. However, the massive variance presents a challenge in running the actor-critic algorithm to estimate $\theta^*$.

**Q-learning** Recall the solution to the POMDP model in which the optimal policy is a function of an information state. Consider the canonical example in which this is the belief state (the sequence of conditional distributions $\{\Pi_k\}$) evolving on the unit simplex $\mathcal{S}$, and assume that the underlying Markov chain $\Phi$ evolves on a finite set so that the simplex is finite-dimensional.

The Q-function $Q^*\colon \mathcal{S} \times \mathsf{U} \to \mathbb{R}$ is the optimal value function associated with the objective (4). To place the equations in standard form denote $c(x,u) = (1-u)c_\circ(x) + uc_\bullet(x)$ for $x \in \mathsf{X}$ and $u \in \{0,1\}$ (recall (4)). For any $u$ and $\beta \in \mathcal{S}$ denote $\mathcal{C}(\beta,u) = \sum_x \beta(x)c(x,u)$.

The value $Q^*(\beta,u)$ is defined to be the minimum of $\sum_\phi \beta(\phi)J(\phi, U_0^\infty)$ over all admissible $U_1^\infty$, subject to $(\Pi_0, U_0) = (\beta, u)$. It satisfies the dynamic programming (DP) equation,

$$Q^*(\beta,u) = \mathcal{C}(\beta,u) + \mathsf{E}[\underline{Q}^*(\Pi_{k+1}) \mid \Pi_k = \beta, U_k = u]$$

with $\underline{H}(\beta) = \min\{H(\beta,0), H(\beta,1)\}$ for any function $H\colon \mathcal{S} \times \mathsf{U} \to \mathbb{R}$.

Q-learning algorithms are based on the characterization: $\mathsf{E}[\mathcal{D}_{k+1}^* \mid \mathcal{Y}_k] = 0$ or each $k$ and any adapted input, with $\mathcal{D}_{k+1}^* = -Q^*(\Pi_k, U_k) + c_k + \underline{Q}^*(\Pi_{k+1})$. with $c_k = (1 - U_k)\mathbf{1}\{\tau_{\mathsf{a}} < k\} + \kappa U_k(\tau_{\mathsf{a}} - k)_+$.

This motivates typical Q-learning algorithms.

Given a parameterized family of real-valued functions $\{Q^\theta : \theta \in \mathbb{R}^d\}$ on $\mathsf{S} \times \mathsf{U}$, the goal is to solve the *projected Bellman equation*: $\bar{f}(\theta^*) = 0$ with

$$\bar{f}(\theta) := \mathsf{E}\big[\{-Q^\theta(\mathcal{X}_k, U_k) + c_k + \underline{Q}^\theta(\mathcal{X}_{k+1})\}\zeta_k\big] \quad (19)$$

where $\{\zeta_k\}$ is a $d$-dimensional stochastic process adapted to the observations.

It is typical to take $\zeta_k = \nabla_\theta Q^\theta(\mathcal{X}_k, U_k)\big|_{\theta_k}$, with $\theta_k$ the estimate at iteration $k$. Theory to-date is largely restricted to a linear function class in which $Q^\theta = \theta^\mathsf{T}\psi$ with $\psi\colon \mathsf{S} \times \mathsf{U} \to \mathbb{R}^d$, and in this case $\zeta_k = \psi(\mathcal{X}_k, U_k)$.

Data required in an algorithm is based on successive runs up to time $\tau_{\mathsf{s}}(n)$ for $n \geq 1$, and with $\tau_{\mathsf{s}}(0) := 0$, which results in the observations $\{\mathcal{X}_k^n, \mathcal{X}_{k+1}^n, U_k^n, c(\Phi_k^n, U_k^n)\}$ for $0 \leq k < \tau_{\mathsf{s}}(n)$. We suppress dependency on $n$ by stringing data together, so that for example

$$U_k := U_{k-\tau_{\mathsf{s}}(n-1)}^n \quad \text{for } \tau_{\mathsf{s}}(n-1) \leq k < \tau_{\mathsf{s}}(n) \text{ and } n \geq 1$$

A version of Q-learning is expressed as the recursion

$$\theta_{k+1} = \theta_k + \alpha_{k+1}G_k\zeta_k\mathcal{D}_{k+1}, \qquad\qquad k \geq 0 \quad (20a)$$

$$\mathcal{D}_{k+1} = -Q^{\theta_k}(\mathcal{X}_k, U_k) + c_k + \underline{Q}^{\theta_k}(\mathcal{X}_{k+1}) \quad (20b)$$

where the matrix gain sequence $\{G_k\}$ is a design choice; Zap Q-learning is in some sense *optimal* [14]; This matrix gain was used in [2] for applications to optimal stopping.

We cannot apply [2], based on the elegant algorithm of [21], since the resulting policy will depend on the cost $\{c(\Phi_k, U_k)\}$ (assumed observed in this prior work).

While there is great empirical success in the history of Q-learning, to-date we only have general conditions for stability of the algorithm, and existence of a solution to the projected Bellman equation [15]. Most crucial is the requirement that the input used for training is an $\varepsilon$-greedy policy (or a smoothed variant). It is shown that, subject to a mild full rank condition for $\psi$, that for sufficiently small $\varepsilon > 0$ the algorithm (20a) is stable in the sense of ultimate boundedness, and there exists at least one solution $\theta^* \in \mathbb{R}^d$ to the projected Bellman equation $\bar{f}(\theta^*) = 0$. Convergence remains a topic of research.

Stability of Zap Q-learning with an oblivious policy (independent of parameter) is virtually universal [3], but this paper makes no claims of existence of $\theta^*$ in this setting. It

is very likely that the main result of [3] can be extended to $\varepsilon$-greedy policies.

## IV. NUMERICAL RESULTS WITH Q-LEARNING

This section contains two subsections: the first illustrates the large-$\kappa$ approximations of Prop. 2.2, and the second summarizes results obtained using the Q-learning formulations described in Section III. Extended details for this section are provided in the preprint version of this paper in [6].

**QCD model**  The conditional i.i.d. model (5) was used to generate observations, in which $X_k^0 \sim N(0, \sigma^2) = f_0$ and $X_k^1 \sim N(\mu_1, \sigma^2) = f_1$ with $\mu_1 = 0.5$ and $\sigma = 1$.

Three choices of $F$ were tested, each of the form $F = \log(\breve{f}_1/\breve{f}_0)$:
**Case 1:** The ideal Gaussian case, in which $\breve{f}_1 = f_1$ and $\breve{f}_0 = f_0$.
**Case 2:** $\breve{f}_0$ is Laplace$(0, b)$ and $\breve{f}_1$ Laplace$(\mu_1, b)$ with $\mu_1 = 0.5$ and $b = \sqrt{\sigma^2/2}$ (matching second order statistics).

Fig. 4.   Three LLRs

**Case 3:** $\breve{f}_0$ is Cauchy$(0, \gamma)$ and $\breve{f}_1$ Cauchy$(x_1, \gamma)$ with $x_1 = 0.5$ and $\gamma$ chosen so that the Gaussian and Cauchy cdfs evaluated at $\sigma = 1$ are equal.

Plots of the three functions are shown in Fig. 4, and the corresponding log moment generating functions are shown in Fig. 3.
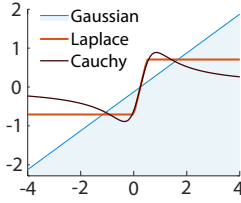
### A. Cost approximation

For the ideal case $F$ is the true LLR for the marginals of $X^0, X^1$, giving $F(x) = L(x) = \mu_1 x - \mu_1^2/2$ and $\Lambda_0(\upsilon) = m_1 \upsilon(\upsilon - 1)$. Hence the equation $\varrho_{\mathsf{a}} = \Lambda_0(\upsilon_+)$ is easily inverted to obtain $\upsilon_+ > 0$ (recall this is required in (10)). For the mismatched case, $\upsilon_+$ was approximated based on numerical computation of $\Lambda_0$.

For each choice of $F$ we obtained estimates via Monte-Carlo of the optimal threshold and corresponding total cost defined in (9), denoted $\{\widehat{\mathrm{H}}^*(\kappa), \widehat{J}^*(\kappa)\}$. In the plots described next the shifted values are displayed:

$$\bar{\mathrm{H}}_{\mathsf{s}}(\kappa) = \bar{\mathrm{H}}_\infty^*(\kappa) - \bar{\mathrm{H}}_\infty^*(100) + \widehat{\mathrm{H}}^*(100) \qquad (21\text{a})$$

$$\bar{J}_{\mathsf{s}}(\kappa) = \bar{J}_\infty^*(\kappa) - \bar{J}_\infty^*(100) + \widehat{J}^*(100) \qquad (21\text{b})$$

This ensures that the approximations (10) coincide with the Monte-Carlo estimates of (9) at $\kappa = 100$.

Fig. 1 shows the Monte Carlo estimates with $1\sigma$ confidence intervals compared to results from experiments in all three cases, and two different choices for the change time, each satisfying $\varrho_{\mathsf{a}} = 0.02$. In the first column the change time has geometric distribution with parameter $\varrho_{\mathsf{a}}$, expressed $\tau_{\mathsf{a}} \sim \mathrm{geo}(\varrho_{\mathsf{a}})$. In the second column the distribution is a mixture of geometrics: $\tau_{\mathsf{a}} \sim \mathrm{geo}(\varrho_{\mathsf{a}})$ with probability 0.05; else $\tau_{\mathsf{a}} \sim \mathrm{geo}(0.2)$.

The approximations are remarkably accurate in all cases, which means that the error $\bar{\mathrm{H}}_\infty^*(\kappa) - \mathrm{H}^*(\kappa)$ is nearly constant over the entire range. Unfortunately, the constant value is large, which motivates learning techniques to obtain a near-optimal threshold.

### B. Q-learning

The remainder of this section presents the design and evaluation of Q-learning for this Bayesian QCD problem. In each experiments the observations come from Shiryaev's model in which the change time is geometrically distributed, and hence the true optimal test is available.

**Basis selection**  The basis for the function class $\{Q^\theta = \theta^\intercal \psi\}$ took the form $\psi(x, u) = (1 - u)\psi^0(x) + u\psi^1(x)$ (recall Section III). A four dimensional basis gave good results:

$$\psi^0(x) = [x; q(x); 0; 0; 0]$$
$$\psi^1(x) = [0; 0; 1; x; q(x))] \qquad (22)$$

with $q(x) = x \exp(-x/b_q)$ for a choice of constant $b_q$.

This basis was chosen based on preliminary experiments with a particular choice of binning: $\psi_i(x, u) = \mathbf{1}\{x \in S_{k_i}, u = u^i\}$ for a collection of intervals $\{S_j\}$ and input values $\{u^j\}$. Further details are provided below; see in particular Fig. 5 and surrounding discussion.

**Stability of Q-learning**  Recent theory recalled in Section III shows that exploration implies stability of Q-learning under mild assumptions on the basis and the oblivious policy. Crucial for stability is that the exploration gain be sufficiently small. In the experiments surveyed here this was taken to be time varying: $\varepsilon_n = \max\{\varepsilon_f, \varepsilon_0 + (1 - n/n_0)(\varepsilon_f - \varepsilon_0)\}$, defined so that $\varepsilon_n = \varepsilon_f < \varepsilon_0$ for $n \geq n_0$. Consistent with theory from [15], it was found in experiments that both the scalar gain algorithm and Zap Q-learning were convergent for sufficiently small $\varepsilon_f$.

In the case of Zap Q-learning, convergence held for a wide range of $\varepsilon_f$; the value $\varepsilon_f = 0.1$ was used in the experiments described in the following. The scalar gain algorithm was *far less reliable*: $\varepsilon_f \leq 10^{-4}$ was required for stability.

One can then ask, is the parameter estimate $\theta^*$ obtained using Zap Q-learning consistent with the output of the scalar gain algorithm? Let $\bar{f}$ denote the vector field for the mean flow associated with the scalar gain Q-learning algorithm—see (19). We computed the Jacobian $\bar{A}(\theta^*) = \partial_\theta \bar{f}(\theta^*)$ for the parameter $\theta^*$ obtained using Zap Q-learning, and discovered that in most cases it had at least one eigenvalue in the strict right half plane in $\mathbb{C}$. In such cases, $\theta^*$ is not asymptotically stable for the mean flow. Standard stochastic approximation theory implies that the parameter $\theta^*$ would not be found using the Q-learning algorithm (20a) with $G_k = I$.

**Oblivious policy**  Exploration was designed to depend on $\kappa$: at the start of episode $i$, a threshold $\mathrm{H}^{\varepsilon,i}(\kappa)$ was drawn uniformly at random from an interval $[a_\kappa, b_\kappa]$. Then, $U_n = \mathbf{1}\{\mathcal{X}_n \geq \mathrm{H}^{\kappa,i}\}$ for each $n$ in this episode.

The threshold approximations of eq. (10) motivated the design of the interval:

$$[a_\kappa, b_\kappa] = [\bar{\mathrm{H}}_\infty^*(\kappa) + \eta - \delta, \bar{\mathrm{H}}_\infty^*(\kappa) + \eta + \delta] \qquad (23)$$

with $\delta > \eta > 0$ constant. This ensured significant exploration in all cases considered.

**Numerical experiments**  Each parameter $\theta \in \mathbb{R}^d$ defines a policy, $\phi^\theta(x) = \arg\min_u Q^\theta(x, u)$ for $x \in \mathbb{R}_+$. In the
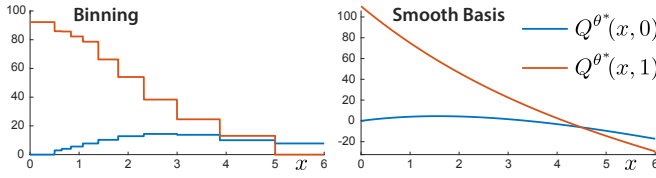
Fig. 5. Insights from binning led to the basis in (22)



Fig. 7. Histograms of $\{Z_1^i : 1 \leq i \leq M\}$ for three values of $N$.

applications considered here this becomes

$$\phi^\theta(x) = \mathbf{1}\{Q^\theta(x,0) \geq Q^\theta(x,1)\}, \qquad x \in \mathbb{R}_+. \quad (24)$$

In every successful application of Q-learning it was found that this policy had a threshold form

$$\phi^\theta(x) = \mathbf{1}\{x \geq \mathrm{H}^\theta\}, \quad \mathrm{H}^\theta > 0 \quad (25)$$

Algorithm performance is investigated in the remainder of this section. For each algorithm, PR-averaging was used to define the final estimate $\hat{\theta}$, and from this a final policy $\widehat{\phi} := \phi^{\hat{\theta}}$ whose performance is compared to the optimal.



Fig. 6. Average cost comparisons.

Initial experiments involved a choice of binning, resulting in $d = 2(d_0 - 1)$, with $d_0$ the number of bins. However, binning proved insufficient for obtaining thresholds close to optimal over all $\kappa$, due in part to a dependence on the choice of bin spacing. This shortcoming is illustrated in Fig. 5, where bin spacing influences the intersection $Q^\theta(x,0) = Q^\theta(x,1)$ for the policy (24). This inspired the "smooth" basis in (22) for which we observed two advantages compared to binning: 1/ better performance of $\widehat{\phi}$ for all $\kappa$, and 2/ reduced training time.

Histograms were generated to evaluate the variance of the parameter estimates. Let $\xi = \xi(N) \geq N$ denote the total number of samples $(\mathcal{X}_k, U_k)$ collected over $N$ episodes, so that $\hat{\theta} = \theta_\xi^{\mathrm{PR}}$ is the final estimate.

The *asymptotic covariance* of $\tilde{\theta}_\xi^{\mathrm{PR}} := \theta_\xi^{\mathrm{PR}} - \theta^*$ is denoted

$$\Sigma_\Theta^{\mathrm{PR}} = \lim_{N \to \infty} \xi \mathsf{E}[\tilde{\theta}_\xi^{\mathrm{PR}}\{\tilde{\theta}_\xi^{\mathrm{PR}}\}^\mathsf{T}]$$

This was estimated using the *batch means method*: $M$ independent runs resulted in the estimates $\{\theta_{\xi^i}^{\mathrm{PR}}, \xi^i : 1 \leq i \leq M\}$. The empirical covariance of $\{Z^i = \sqrt{\xi^i}[\theta_{\xi^i}^{\mathrm{PR}} - \bar{\theta}^{\mathrm{PR}}] : 1 \leq i \leq M\}$ provides an estimate of $\Sigma_\Theta^{\mathrm{PR}}$ with vanishing error as $M \to \infty$ and then $N \to \infty$.

An example is shown in Fig. 7 for the case $\kappa = 27$, using $M = 400$ and three different values of $N$. Only the fourth component of the five dimensional histogram is shown, giving an estimate of $\Sigma_\Theta^{\mathrm{PR}}(4,4)$. What is crucial here is that the estimate of this value is nearly identical for the three values of $N$ chosen. Similar results were observed for estimates the other diagonal entries of $\Sigma_\Theta^{\mathrm{PR}}$.
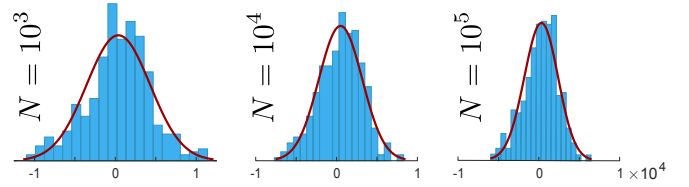
This is an example of how the CLT can be used to estimate required run lengths by first conducting a large number of independent experiments with a relatively short run length—in this example, $N = 10^4$ provides a reasonable estimate of the variance of $Z^i = \sqrt{\xi^i}[\theta_{\xi^i}^{\mathrm{PR}} - \bar{\theta}^{\mathrm{PR}}]$ for each $i$ and $N \gg 10^4$.

Thresholds $\mathrm{H}_N^{\theta,i}$ yielded smaller empirical variance than $\theta_N^i$ for all $\kappa$. Example histograms showing the fourth parameter $\theta_N^i(4)$ are included in Fig. 8 for $M = 30$ and $N = 10^6$.
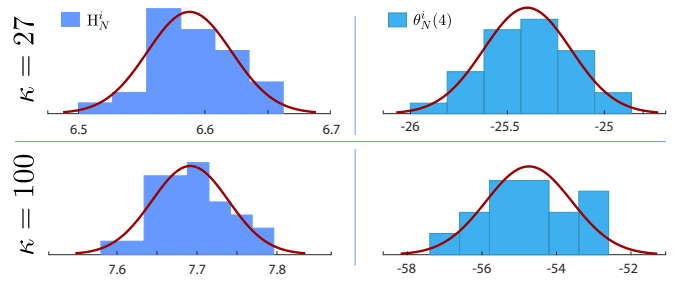


Fig. 8. Histograms of $\{\mathrm{H}_N^{\theta,i}, \theta_N^i(4) : 1 \leq i \leq M\}$ for $N = 10^6$

Fig. 6 shows the average cost of the policy $\widehat{\phi}$ for the ideal Gaussian case. Also included are estimates $\widehat{J}^*(\kappa)$ using CUSUM* and the *optimal test*—Shiryaev test defined below (5). The confidence intervals for each are of $1\sigma$ standard deviation. One metric for success is that the shape of the average cost curve obtained through Q-learning resembles the optimal, which we observed for $\kappa \geq 20$.

*Mismatched cases.* Experiments were repeated for Cases 2 and 3: the surrogate information state $\{\mathcal{X}_n\}$ differs in each case based on the respective LLRs, plotted in Fig. 4.

Recall the discussion surrounding (24): the policy $\phi^\theta$ obtained from Q-learning resulted in a threshold policy in almost all cases. Therefore, the best performance from Q-learning can be no better than CUSUM*, which uses the optimal threshold (9a). Moreover, the performance in Cases 2 and 3 is poor when compared with Case 1—recall the comparisons in Fig. 1.

However, in all three cases, the policy obtained from Q-learning using $\kappa \geq 27$ yielded average cost within 5% of its respective CUSUM* cost estimate $\widehat{J}^*(\kappa)$.

## V. CONCLUSIONS

The theory and numerical results in this paper motivate many directions for future research:

• The preliminary research surveyed in Section III motivates research on variance reduction for stochastic gradient descent.

- The independence assumptions in Prop. 2.2 are unfortunate. Relaxation of the i.i.d. assumption is the main topic of [4], [5], along with conditions under which the independence of $\tau_{\mathsf{a}}$ and $X^0$ may be relaxed; the assumptions of Lemma 2.1 are sufficient to obtain an extension of Prop. 2.2.

- In applications of interest to us there may be well understood behavior before a change (which might represent a fault in a transmission line, or a computer attack). We cannot expect to have a full understanding of post-change behavior. The choice of surrogate information state must be reconsidered in these settings, perhaps based on techniques for universal hypothesis testing (see [22], [8] and the references therein).

## VI. Appendix

Below we include simulation details for the QCD estimates and RL experiments.

**Actor-critic method** Estimates of $\Sigma_{\nabla}$ were obtained by averaging $N = 10^7$ independent episodes.

Estimates of the gradient $\breve{\nabla}_{\Gamma}(\theta)$ were obtained using a much shorter run, with $N = 10^4$. Two estimates of the objective were produced with $N = 10^4$: $\Gamma(\theta)$ using Monte-Carlo to estimate the expectation in (12) directly and $\widehat{\Gamma}(\theta)$ through numerical integration.

**Q-learning for QCD** For the ideal and mismatched cases, Monte Carlo simulations were used to estimate MDE and MDD for CUSUM*. Estimates were also obtained for the optimal Shiryaev test $\mathcal{X}_n = p_n = \mathsf{P}\{\tau_{\mathsf{a}} \leq n \mid Y_0^n\}$. Parameters for the stochastic processes $\{X_k^0, X_k^1\}$ and $\tau_{\mathsf{a}}$ matched those used for Q-learning. For both simulations, $N = 2e4$ sample paths were run. A range of $T = 10^3$ thresholds $0 \leq \mathsf{H} \leq 20$ was used for CUSUM*, and $0 \leq \mathsf{H} \leq 1$ for Shiryaev. For each H, a pair MDE(H) and MDD(H) was obtained by averaging the eagerness and delay over $N$ runs. This repeated for $M = 200$ independent runs, averaging again to obtain for each test a $T \times 2$ matrix [MDE, MDD], where each row corresponds to a different threshold H. Estimates of these quantities are random variables, whose variances were found to be very small. Estimates $\widehat{\mathsf{H}}^*(\kappa)$ and $\widehat{J}^*(\kappa)$ as described before eq. (21) were then obtained for a range $2 \leq \kappa \leq 100$ to generate the average curves in Fig. 6.

## References

[1] V. Borkar, S. Chen, A. Devraj, I. Kontoyiannis, and S. Meyn. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *arXiv e-prints:2110.14427*, pages 1–50, 2021.

[2] S. Chen, A. M. Devraj, A. Bušić, and S. Meyn. Zap Q-learning for optimal stopping. In *Proc. of the American Control Conf.*, pages 3920–3925, 2020.

[3] S. Chen, A. M. Devraj, F. Lu, A. Bušić, and S. Meyn. Zap Q-Learning with nonlinear function approximation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Proc. Conference on Neural Information Processing Systems (NeurIPS), and arXiv e-prints 1910.05405*, volume 33, pages 16879–16890, 2020.

[4] A. Cooper and S. Meyn. Quickest change detection using mismatched CUSUM. *arXiv 2409.07948*, 2024.

[5] A. Cooper and S. Meyn. Quickest change detection using mismatched CUSUM—extended abstract. In *Allerton Conference on Communication, Control, and Computing*, 2024.

[6] A. Cooper and S. Meyn. Reinforcement learning design for quickest change detection—extended paper. *arXiv preprint arXiv:2403.14109*, 2024.

[7] R. J. Elliott, L. Aggoun, and J. B. Moore. *Hidden Markov models*, volume 29 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1995. Estimation and control.

[8] D. Huang and S. Meyn. Generalized error exponents for small sample universal hypothesis testing. *IEEE Trans. Inform. Theory*, 59(12):8157–8181, 2013.

[9] V. Krishnamurthy. Structural results for partially observed Markov decision processes. *ArXiv e-prints*, page arXiv:1512.03873, 2015.

[10] M. N. Kurt, O. Ogundijo, C. Li, and X. Wang. Online cyber-attack detection in smart grid: A reinforcement learning approach. *IEEE Transactions on Smart Grid*, 10(5):5174–5185, 2019.

[11] Y. Li, C. Szepesvari, and D. Schuurmans. Learning exercise policies for American options. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 352–359, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[12] Y. Liang, A. G. Tartakovsky, and V. V. Veeravalli. Quickest change detection with non-stationary post-change observations. *arXiv 2110.01581*, 2021.

[13] Y. Liang and V. V. Veeravalli. Non-parametric quickest mean-change detection. *Transactions on Information Theory*, pages 8040–8052, 2022.

[14] S. Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, Cambridge, 2022.

[15] S. Meyn. Stability of Q-learning through design and optimism. *arXiv 2307.02632*, 2023.

[16] G. V. Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379 – 1387, 1986.

[17] M. Pollak and A. G. Tartakovsky. On optimality properties of the Shiryaev-Roberts procedure. *Statistica Sinica*, 19(4):1729–1739, 2009.

[18] A. S. Polunchenko and A. G. Tartakovsky. On optimality of the Shiryaev–Roberts procedure for detecting a change in distribution. *The Annals of Statistics*, 38(6):3445 – 3457, 2010.

[19] A. N. Shiryaev. *Optimal stopping rules*, volume 8. Springer Science & Business Media, 2007 (reprint from 1977 ed.).

[20] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *The Journal of Machine Learning Research*, 23(1):483–565, 2022.

[21] J. Tsitsiklis and B. van Roy. Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Trans. Automat. Control*, 44(10):1840 –1851, 1999.

[22] J. Unnikrishnan, D. Huang, S. P. Meyn, A. Surana, and V. V. Veeravalli. Universal and composite hypothesis testing via mismatched divergence. *IEEE Trans. Inform. Theory*, 57(3):1587 –1603, 2011.

[23] V. V. Veeravalli and T. Banerjee. Quickest change detection. In *Academic press library in signal processing*, volume 3, pages 209–255. Elsevier, 2014.

[24] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli. Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2):494–514, 2021.

[25] Q. Zhang, Z. Sun, L. C. Herrera, and S. Zou. Data-driven quickest change detection in hidden Markov models. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2643–2648, June 2023.

[26] Q. Zhang, Z. Sun, L. C. Herrera, and S. Zou. Data-driven quickest change detection in Markov models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, June 2023.

[27] E. Zhou. Optimal stopping under partial observation: Near-value iteration. *IEEE Transactions on Automatic Control*, 58(2):500–506, 2013.