Minimum-Delay Adaptation in Non-Stationary Reinforcement Learning via Online High-Confidence Change-Point Detection

Lucas N. Alegre [1] **Ana L. C. Bazzan** [1] **Bruno C. da Silva** [2]

Abstract

Non-stationary environments are challenging for reinforcement learning algorithms. If the state transition and/or reward functions change based on latent factors, the agent is effectively tasked with optimizing a behavior that maximizes performance over a possibly infinite random sequence of Markov Decision Processes (MDPs), each of which drawn from some unknown distribution. We call each such MDP a *context*. We introduce an algorithm that analyzes a possibly infinite stream of data and computes, in real-time, high-confidence change-point detection statistics that reflect whether novel, specialized policies need to be created and deployed to tackle novel contexts, or whether previously-optimized ones might be reused. We show that *(i)* this algorithm minimizes the delay until unforeseen changes to a context are detected, thereby allowing for rapid responses; and *(ii)* it bounds the rate of false alarm, which is important in order to minimize regret. [1]

1. Introduction

Reinforcement learning (RL) techniques have been successfully applied to solve high-dimensional sequential decision problems. However, if the state transition and/or reward functions change unexpectedly, according to latent factors unobservable to the agent, the system is effectively tasked with optimizing behavior policies that maximize performance over a (possibly infinite) random sequence of Markov Decision Processes (MDPs). Each MDP is drawn from an unknown distribution and is henceforth referred to as a *context*. Designing efficient algorithms to tackle this problem is a known challenge in RL (Padakandla, 2020). The key difficulties here result from *(i)* the need to quickly and reliably detect when the underlying system dynamics has changed; and *(ii)* the need to effectively learn and deploy adaptable prediction models and policies, specialized in particular contexts, while allowing the agent to (when appropriate) reuse previously-acquired knowledge.

Many existing related works tackle non-stationary problems either by detecting when the underlying MDP changes, or via meta-learning approaches that construct a prior model (or policy) capable of rapidly generalizing to novel contexts. Hadoux et al., for example, introduced a technique based on change-point detection algorithms to deal with non-stationary problems with discrete state spaces (Hadoux et al., 2014; Banerjee et al., 2017). We, by contrast, address the more general setting of high-dimensional continuous RL problems. Supervised meta-learning algorithms (Finn et al., 2017) have also been recently combined with RL to enable fast adaptation under changing domains (Nagabandi et al., 2019a;b). Meta-learning methods typically assume that the distribution over contexts experienced during training is the same as the one experienced during testing. We, by contrast, do not require that contexts are sampled from a previously-seen distribution, nor that contexts share structural similarities with previously-experienced ones.

To address these limitations, we introduce an algorithm that analyzes a possibly infinite stream of data and computes, in real-time, high-confidence change-point detection statistics that reflect whether novel, specialized policies need to be deployed to tackle new contexts, or whether a previously-optimized policy may be reused. We call our algorithm Model-Based RL Context Detection, or MBCD. We formally show that it minimizes the delay until unforeseen changes to a context are detected, thereby allowing for rapid responses, and that it allows for formal bounds on the rate of false alarm—which is of interest when minimizing the agent's regret over random sequences of contexts. Our method constructs a mixture model composed of a (possibly infinite) ensemble of probabilistic dynamics predictors that model the different modes of the distribution over underlying latent MDPs.

[1] Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil [2] CICS, University of Massachusetts, Amherst, USA. Correspondence to: Lucas N. Alegre <lnalegre@inf.ufrgs.br>.

Proceedings of the LatinX in AI (LXAI) Research workshop at ICML 2021. Copyright 2021 by the author(s).

[1] A full-version of this paper is published in the Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS) , 2021.

2. Problem Formulation

We define a non-stationary environment as a family of MDPs $\{\mathcal{M}_z\}_{z \in \mathbb{N}^+}$. Each MDP $\mathcal{M}_z$ is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}_z, \mathcal{R}_z, \gamma, d^0)$, where $\mathcal{S}$ is a (possibly continuous) state space, $\mathcal{A}$ is a (possibly continuous) action space, $\mathcal{T}_z : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a transition function specifying the distribution over next states, given the current state and action, $\mathcal{R}_z : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function, $\gamma \in [0, 1]$ is the discount factor, and d^0 is an initial state distribution. In what follows, S_t, A_t, and R_t are the random variables corresponding to the state, action, and reward at time step t. We assume that the agent observes a random sequence $(\mathcal{M}_0, \mathcal{M}_1, \ldots)$ of MDPs—called *contexts*—drawn independently from some unknown distribution. We assume that the number of contexts, $|\{\mathcal{M}_z\}|$, is unknown. Let z be a latent index variable indicating a particular MDP, $\mathcal{M}_z$. We assume that each MDP's transition and reward function are parameterized by a latent vector θ_z. Let $p_{\theta_z}(S_{t+1}, R_t | S_t, A_t)$ denote the joint conditional probability distribution over next-state and reward associated with MDP $\mathcal{M}_z$. We do not impose any smoothness assumptions on how variations to θ_z affect $\mathcal{T}_z$ and $\mathcal{R}_z$: contexts may be arbitrarily different and share no structural similarities.

Let the time steps in which context changes occur be an increasing sequence of integer random variables, $\{C_i\}_{i \geq 1}$, for which a prior $\phi(C_i)$ is unknown or cannot be defined. We call each C_i a *change-point*. At every change-point C_i, the current context $\mathcal{M}_z$ is replaced by a new randomly drawn MDP. To perform well, an agent must rapidly detect context changes and deploy an appropriate policy. If a new random context differs significantly from previously-experienced ones, the agent may have to learn a policy from scratch; otherwise, it may choose to reuse previously-acquired knowledge to accelerate learning and avoid catastrophic forgetting.

3. High-Confidence Change-Point Detection

In the *online* CPD setting, a sequential detection procedure is defined to rapidly and reliably estimate when the parameter θ of some underlying distribution or stochastic process has changed. Online CPD algorithms should produce high-confidence estimates, Γ, of the true change-point time, C. Notice that Γ is a random variable whose stochasticity results from the unknown stochastic prior over context changes, ϕ, and from the fact that each MDP in $\{\mathcal{M}_z\}$ produces random trajectories of states, actions, and rewards.

Suppose that at each time t, while the agent interacts with $\mathcal{M}_0$, sample next-state and rewards are drawn from $p_{\theta_0}(S_{t+1}, R_t | S_t, A_t)$, where θ_0 is the latent vector parameterizing $\mathcal{M}_0$'s transition and reward functions. At some unknown random change-point C, the context changes to $\mathcal{M}_1$, and experiences that follow are drawn from $p_{\theta_1}(S_{t+1}, R_t | S_t, A_t)$. We propose to identify such a change by computing high-confidence statistics that reflect whether θ_0 has changed. This can be achieved by introducing a minimax formulation of the CPD problem, as discussed by (Pollak, 1985). In this formulation, the goal is to minimize the worst-case expected detection delay, $\Delta_{worst}(\Gamma)$, associated with the random estimates Γ produced by a particular CPD algorithm (when considering all possible change-points C), given that a bound on the maximum false alarm rate (FAR) may be imposed. The worst-case expected detection delay, $\Delta_{worst}(\Gamma)$, and the FAR, are defined as:

$$\Delta_{worst}(\Gamma) = \sup_{c \geq 1} \, \mathbb{E}[\Gamma - C | \Gamma \geq C, C = c], \quad (1)$$

$$\text{FAR}(\Gamma) = \frac{1}{\mathbb{E}[\Gamma | C = \infty]}, \quad (2)$$

where the expectations in Eq. 1 and Eq. 2 are over the possible histories of experiences produced by the stochastic process, and where conditioning on $C = \infty$ indicates the random event where the context never changes. Given these definitions, the objective of a high-confidence change-point detection process is the following:

$$\inf_{\Gamma} \Delta_{worst}(\Gamma) \text{ subject to } \text{FAR}(\Gamma) \leq \alpha, \quad (3)$$

where α denotes the desired upper-bound on the false alarm rate.

4. Model-Based RL Context Detection

In this section, we introduce an algorithm that iteratively applies a CUSUM-related procedure to detect context changes under the assumptions discussed in Section 2. The algorithm incrementally builds a library of models and policies for tackling arbitrarily different types of contexts; i.e., contexts that may result from quantitatively and qualitatively different underlying causes for non-stationarity—ranging from unpredictable environmental changes (such as random wind) to robot malfunctions. Our method can rapidly deploy previously-constructed policies whenever contexts approximately re-occur, or learn new decision-making strategies whenever novel contexts, with no structural similarities with respect to previously-observed ones, are first encountered.

We now introduce a *high-level* description of our method (Model-Based RL Context Detection, or MBCD). As the agent interacts with a non-stationary environment, context changes are identified via a multivariate variant of CUSUM (Healy, 1987), called MCUSUM. These statistics inherit the same formal properties as those presented in Section 3. In particular, they formally guarantee that MBCD can detect context changes with minimum expected delay, while simultaneously bounding the false alarm rate. As a consequence, MBCD can effectively identify novel contexts while ensuring, with high probability, that new context-specific policies will only be constructed when necessary.

As new contexts are identified by this procedure, MBCD updates a mixture model, M, composed of a (possibly infinite) ensemble of probabilistic context dynamics predictors, whose purpose is to model the different modes of the distribution over underlying latent MDPs/contexts. New models are added to the ensemble as qualitatively different contexts are first encountered. The mixture model M associates, with each identified context $\mathcal{M}_z$, a learned joint distribution p_{θ_z} over next-state dynamics and rewards associated. Let K be the number of context models currently in the mixture. After each agent experience, MBCD identifies the most likely context, z_t, by analyzing a set of incrementally-estimated MCUSUM statistics. Whenever a novel context—one with dynamics that are qualitatively different from those previously-experienced—is observed, a new model is added to the mixture. Context-specific policies, π_{ψ_z}, are trained via a Dyna-style approach (Sutton, 1990) based on the corresponding learned joint prediction model of $\mathcal{M}_z$, p_{θ_z}.

4.1. Online Context Change-Point Detection

In the particular case where the dynamics of each context are modeled as multivariate Gaussians, the Log-Likelihood Ratio (LLR) statistic can be computed as follows. To simplify notation, let $\mu_0 = \mu_{\theta_0}(S_t, A_t)$ and $\Sigma_0 = \Sigma_{\theta_0}(S_t, A_t)$. It is then possible to show that the LLR statistic, L_t, between distributions $p_{\theta_1}(\mathbf{Y}_t|\mathbf{X}_t)$ and $p_{\theta_0}(\mathbf{Y}_t|\mathbf{X}_t)$, is given by:

$$L_t = \log \frac{(2\pi)^{-\frac{d}{2}}|\Sigma_1|^{-\frac{1}{2}}\exp\{-0.5(\mathbf{Y}_t-\mu_1)\Sigma_1^{-1}(\mathbf{Y}_t-\mu_1)\}}{(2\pi)^{-\frac{d}{2}}|\Sigma_0|^{-\frac{1}{2}}\exp\{-0.5(\mathbf{Y}_t-\mu_0)\Sigma_0^{-1}(\mathbf{Y}_t-\mu_0)\}} \tag{4}$$

where d is the dimensionality of the multivariate Gaussian. At each time step t, MBCD uses L_t to compute MCUSUM statistics $W_{k,t}$ for each known context k:

$$W_{k,t} \leftarrow \max\left(0, W_{k,t-1} + \log\frac{p_{\theta_k}(\mathbf{Y}_t|\mathbf{X}_t)}{p_{\theta_{z_t}}(\mathbf{Y}_t|\mathbf{X}_t)}\right), \ \forall k \in [1, K]. \tag{5}$$

Given updated statistics $W_{k,t}$ and a decision threshold h, the most likely current context, z_t (which may or may not have changed) can then be identified as:

$$z_t \leftarrow \begin{cases} \text{argmax}_k\, W_{k,t}, & \text{if } \exists k \in [1, K] \cup [\text{new}] \text{ s.t. } W_{k,t} > h, \\ z_{t-1}, & \text{otherwise.} \end{cases} \tag{6}$$

If no alternative contexts are more likely to have generated the observations collected up to time t, no context change is detected and $z_t = z_{t-1}$.

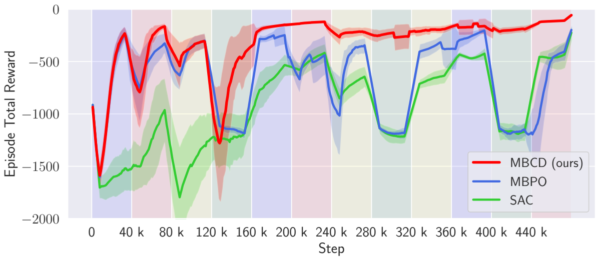

Figure 1. Evaluation of MBCD on the non-stationary *Half-Cheetah* domain. Colored shaded areas represent different contexts: *(blue)* default context; *(red)* joint malfunction; *(yellow)* wind; *(green)* novel target velocity.

5. Experiments

We evaluate our method on the non-stationary Half-Cheetah domain and compare it with two state-of-the-art RL algorithms: MBPO (Janner et al., 2019) and SAC (Haarnoja et al., 2018). In our setting, MBPO can be seen as a particular case of our algorithm, where a single dynamics model and policy are tasked with optimizing behavior under changing contexts. SAC works similarly to MBPO but does not perform Dyna-style planning steps using a learned model.

Fig. 1 shows the total reward achieved by different methods (ours, MBPO, SAC) as contexts change. Colored shaded areas depict different contexts, as discussed in the figure's caption. Notice that our method and MBPO have similar performances when interacting for the first time with the first three random contexts. In particular, both MBCD and MBPO's performances temporarily drop when a novel context is encountered for the first time. MBCD's performance drops because it instantiates a new dynamics model for the newly encountered context, while MBPO's performance drops because it undergoes negative transfer. SAC, which is model-free, never manages to achieve reasonable performance during the duration of each context, due to sample inefficiency. However, as the agent encounters contexts with structural similarities with respect to previously-encounters ones (around time step 160k), MBCD's performance becomes near-optimal: it rapidly identifies whenever a context change has occurred and deploys an appropriate policy. MBPO and SAC, on the other hand, suffer from negative transfer due to learning average policies or dynamics models. They are also subject to catastrophic forgetting and do not reuse previously-acquired, context-specific knowledge.

Next, we analyze how MBCD performs when compared with state-of-the-art meta-learning methods specifically tailored to deal with non-stationary environments: Gradient-Based Adaptive Learner[2] (GrBAL) (Nagabandi et al., 2019a)

[2] We used the authors' implementation of the method,

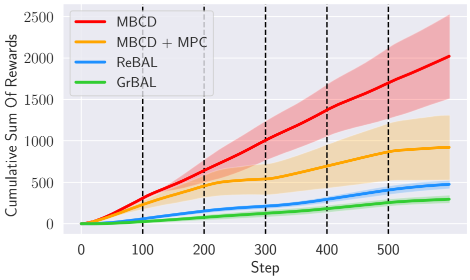

Figure 2. Performance of MBCD and meta-learning methods (after a pre-training phase) in the *Half-Cheetah* domain with non-stationary malfunctions that disable random joints. Vertical dashed lines indicate context changes.

and Recurrence-Based Adaptive Learner[2] (ReBAL) (Nagabandi et al., 2019a). Fig. 2 compares the adaptation performance of MBCD and the meta-learning methods in a non-stationary setting where (inspired by (Nagabandi et al., 2019a)) random joints of the Half-Cheetah robot are disabled after every 100 time steps. In this experiment we compare MBCD, ReBAL, GrBAL, and also (for fairness) a variant of MBCD that chooses actions using MPC instead of SAC. Although the meta-learning methods have lower-variance, their meta-prior models do not perform as well as the MBCD context-specific dynamics models and policies. We also observe that when MBCD uses parameterized policies, learned through Dyna-style planning, it performs better than MBCD coupled with MPC.

6. Conclusion

We introduced a model-based reinforcement learning algorithm (MBCD) that learns efficiently in non-stationary settings with continuous states and actions. It makes use of high-confidence change-point detection statistics to detect context changes with minimum delay, while bounding the rate of false alarm. It is capable of optimizing policies online, without requiring a pre-training phase, even when faced with streams of arbitrarily different contexts drawn from unknown distributions. We empirically show that it outperforms state-of-the-art (model-free and model-based) RL algorithms, and that it outperforms state-of-the-art meta-learning methods specially designed to deal with non-stationarity

References

Banerjee, T., Liu, M., and How, J. P. Quickest change detection approach to optimal control in markov decision processes with model changes. In *2017 American Control Conference (ACC)*, pp. 399–405, Seattle, WA, USA, 2017. IEEE. ISBN 150905992X.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135, Sydney, Australia, August 2017. PMLR.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870, Stockholmsmässan, Stockholm Sweden, July 2018. PMLR.

Hadoux, E., Beynier, A., and Weng, P. Sequential decision-making under non-stationary environments via sequential change-point detection. In *Learning over Multiple Contexts (LMCE)*, Nancy, France, September 2014.

Healy, J. D. A note on multivariate cusum procedures. *Technometrics*, 29(4):409–412, 1987. ISSN 00401706.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems (NIPS) 32*, pp. 12519–12530, Vancouver, Canada, 2019. Curran Associates, Inc.

Nagabandi, A., Clavera, I., Liu, S., Fearing, R. S., Abbeel, P., Levine, S., and Finn, C. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, May 2019a.

Nagabandi, A., Finn, C., and Levine, S. Deep online learning via meta-learning: Continual adaptation for model-based rl. In *7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, May 2019b.

Padakandla, S. A survey of reinforcement learning algorithms for dynamically varying environments. *arXiv e-prints*, May 2020.

Pollak, M. Optimal detection of a change in distribution. *The Annals of Statistics*, 13:206–227, March 1985.

Sutton, R. S. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the 7th International Conference on Machine Learning (ICML)*, pp. 216–224, Austin, TX, USA, 1990.

publicly available at `https://github.com/iclavera/learning_to_adapt`.