ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9.
Available from http://www.i6doc.com/en/.

Robust and Cheap Safety Measure for Exoskeletal Learning Control with Estimated Uniform PAC (EUPAC)

Felix Weiske and Jens Jäkel

University of Applied Sciences Leipzig - Faculty of Engineering
Karl-Liebknecht-Straße 132, 04277 Leipzig - Germany

Abstract. Although safe reinforcement learning control for exoskeletons shows great potential, established real-world applications seem rare. There is a dilemma: the safe RL agent is either robustly safe and computationally demanding or not robustly safe but computationally cheap. We propose Estimated Uniform PAC (EUPAC) as a new safety heuristic. We show that our EUPAC algorithm differentiates safe from unsafe system behaviour with high significance ($p < 0.001$) while having a linear worst time complexity.

1 Introduction

Exoskeletons in close proximity to human users obviously cannot allow for catastrophically unsafe behaviour, which in contrast is somewhat needed for reinforcement learning (RL) agents to explore the learning space properly.

Learning control needs to be safe. Following [1], the underlying methodologies of modern safe learning approaches revolve around changing the optimization criterion and changing the exploration process.

Nowadays, most prominent approaches to changing the optimization criterion are added constraints that ensure safety. There are Control Barrier Functions (CBF) [2, 3] and constraints from Hamilton-Jacobi-Bellman reachability analysis [4]. Though their theoretical guarantees are strong, their computational cost can be high. They restrict the exploration process, need demanding hardware which in turn can be detrimental to the user experience. The latter is particularly relevant for exoskeletal systems, as cost, weight, and comfort are important design considerations. [5].

Changes to the exploration process are made by risk-aware heuristics [6] or by initial or on-the-fly knowledge demonstrations [7, 8]. Risk-aware heuristics are computationally cheap, but cannot guarantee strong confidence bounds around the estimated safety. Knowledge demonstrations introduce data bias, that can reduce learned optimality.

Strong guarantees of safety may be an intuitive demand without question but the cost in performance may also be a crucial hinderance for real-world applications [9]. This is a dilemma: the Safe RL agent can either be safe with high certainty but computationally costly or heuristically safe and computationally cheap.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Can there be a strongly guaranteed (that is, more robust) safety heuristic which is computationally cheap? For this purpose, we derive a numerical estimation of the Uniform PAC (UPAC) learning performance bound [10]. It bounds all possible tolerance errors for regrets happening in a learning setting which is more robust than and even implies PAC [11] and Uniform High Probability Regret bounds [12]. In a preliminary simulative safety benchmark we illustrate Estimated UPAC (EUPAC) by Interval Checking: a heuristic that identifies safe agent behaviour, is more robust, and has cheap computational cost.

2 Uniform PAC-Learning and its estimation

A reinforcement learning problem consists of a (possibly in)finite number of states S, (possibly in)finite number of actions A, time steps k with a total of H per episode. We define regret functions $\Delta : \mathbb{R} \to \mathbb{R}_{\geq 0}$ by reward $\Delta_R(R_k) = R_{\max} - R_k$ with rewards at the current time step R_k and the optimal reward $R_{\max}$, and by observation $\Delta_S(s_k) = \max\{s_k - s_u,\, s_l - s_k,\, 0\}$ with the current state s_k, its safe upper limit s_u and its safe lower limit s_l. Uniform PAC (UPAC)

$$\mathbb{P}\left(\forall \tau > 0 :\ N_\tau(\Omega) \leq F(S, A, H, 1/\tau, \log(1/\delta)) \geq 1 - \delta\right.$$

necessitates the number of τ-regrets $N_\tau(\Omega) = \sum_{\Delta \in \Omega} \mathbb{1}_{\Delta > \tau}$ to be less than a polynomial upper bound F in S, A, H, $1/\tau$ and $\log(1/\delta)$ with probability $1 - \delta$ for all τ [10].

To make UPAC usable as an online heuristic, we derive a numerical estimate. This encompasses 1) linking the UPAC probability to a window of W' seen regrets $\Omega_{W'}$ and their density $\mathbb{P}_\Delta$, and 2) checking if the resulting regret probability fulfills the UPAC condition at the current timestep k. For a detailed derivation, please refer to the supplemental material on the GitHub repository github.com/flxweiske/eupac.

It can be shown that Estimated UPAC (EUPAC)

$$\mathrm{EUPAC}(\Omega) = \sum_k \mathbb{P}(\Omega_{W',k}) \mathbb{P}\left(\forall \tau > 0 :\ N_\tau(\Omega_{W',k}) \leq \max\left(\bar{F}(1/\tau),\, 0\right)\right) \quad (1)$$

is a numerical estimate of UPAC. Since N_τ is at most W', gets reduced by 1 at any $\tau = \Delta_i$ and cannot be lower than 0, it is a descending stair function that allows UPAC to be checked for τ across W' intervals individually. Quantifier elimination reduces the UPAC criterion to a set of interval conditions on EUPAC parameters and seen regrets Δ_i (see Table 1). If

$$\exists \tau : N_\tau > \bar{F} = \bigvee_{i,j}^{W',2} C_i^\tau \wedge C_j^{\bar{F}} \wedge C_{ij}^{\mathrm{EUPAC}}$$

is true, the agent is prone to be more unsafe. Vice-versa, regrets that are safe by the UPAC criterion will add to the EUPAC value by their probability of occurence – the agent is deemed more safe. We calculate EUPAC by Interval

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

$\exists\tau : \neg E$	$C_1^F : \bar{F} = c_1 + c_2/\tau > 0$	$C_2^F : \bar{F} = c_1 + c_2/\tau < 0$
$C_1^\tau : 0 < \tau < \Delta_1$	$C_{11}^{\text{EUPAC}} : W > c_1 + c_2/\tau$	$C_{12}^{\text{EUPAC}} : W > 0$
$C_i^\tau : \Delta_i < \tau < \Delta_{i+1}$	$C_{i1}^{\text{EUPAC}} : W - i > c_1 + c_2/\tau$	$C_{i2}^{\text{EUPAC}} : W - i > 0$
$C_W^\tau : \tau > \Delta_W$	$C_{W1}^{\text{EUPAC}} : 0 > c_1 + c_2/\tau$	$C_{W2}^{\text{EUPAC}} : 0 > 0$

Table 1: Interval conditions that result from the descending stair of N_τ

Checking with Algorithm 1. Our implementation will use multinomial subsets of the seen regrets. To bound the inherent computational burden we actually bin those into W regrets defined by the $W + 2$ boundaries $N_\Delta(\Omega_W) = \bar{F}(1/\Delta)$. For further details, please refer to the supplemental material on the GitHub repository.

Algorithm 1 General steps of EUPAC by Interval Checking

Observe new regret window $\Omega_{W',\,k}$
Bin into regret window with fixed size W
Get a new regret probability estimate $\mathbb{P}_\Delta^k$ for the binned regrets Ω_W
Average with the previous regret probability estimate $\mathbb{P}_\Delta^{k-1}$ via some α
Calculate EUPAC by Interval Checking (1) across all multinomial cases of sample size N_{MN} with the underlying regret probability estimate

3 Safety Benchmark

Main purpose of the benchmark is to show that EUPAC by Interval Checking 1) distinguishes safe and unsafe agent behaviour, 2) is more sensitive to unsafe than safe behaviour and 3) is computationally cheap, that is it has subexponential time complexity.

We studied impedance controlled pendulum system environments between 1 and 3 degrees of freedom that allude to sagittal leg movement models for exoskeletons [13, 14] with suitable parameters [15]. A joint is deemed to be safe if its angle is between the biomechanically feasible thresholds [13].

The reinforcement learning problem learns across all-together 2 million learning steps with 500 timesteps each episode (simulated time of 5 seconds) to achieve a random target position known to the agent. The driving reward-function consists of four components: 1) distance to current target in cartesian coordinates, 2) discrete bonus for high proximity to the target of 100, 3) action regularization with a factor of 0.1 and 4) a duration punishment of -1 at each time step the agent has not achieved high proximity. We use three of the most prominent model-free algorithms DDPG [16], TRPO [17] and SAC [18] to show their individual learning development. To test EUPAC-Safety we use CBF-based compensating control in an end-to-end RL perspective [19] as a wrapper for environments with safe joint dynamics, that is generally safer environments. EUPAC settings can be found in the supplemental material on the GitHub repository.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

	w/o CBF	w/ CBF	stat. diff. (p-Value)
Δ_R	698 ± 978	319 ± 560	no ($p = .0760$)
EUPAC	$64\% \pm 47\%$	$81\% \pm 38\%$	yes ($p = .0115$)
Δ_S	0.31 ± 0.23	0.03 ± 0.07	yes ($p < .0001$)
EUPAC	$33\% \pm 47\%$	$78\% \pm 39\%$	yes ($p < .0001$)

Table 2: Statistical difference for regret and corresponding EUPAC with Mann-Whitney-U because of non-normality of all groupings averaged across all simulations

The benchmark ran on a Intel i7-10870H with 2.2 GHz and a GeForce RTX 3060. Implementations use Python with Numpy, Scipy Optimization, Tensorflow, and Stable Baselines [20]. EUPAC by Interval Checking and all of the supplemental material can be found on the GitHub repository.

4 Results

Table 2 summarizes the results for EUPAC to distinguish safe from unsafe behaviour. The values for EUPAC show statistically significant differences for unsafe and safe behaviour with regrets by reward and regrets by observation even if there is no distinction in the respective regrets.

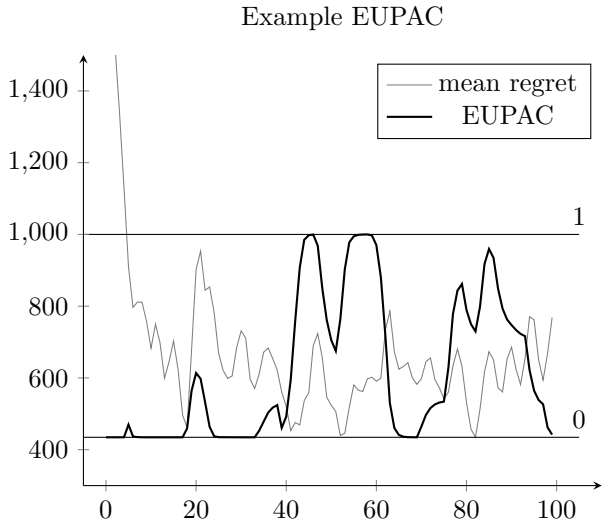

Fig. 1: Typical EUPAC values alongside the regret learning curve

To illustrate the robustness of EUPAC with respect to various regret values, we look at a typical regret learning example (Fig 1). After the initial learning phase between 0% and around 10% of learning episodes, the agent achieves on average lower regret values. Note that as soon as the first regret reaches a

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

N_{MN}	$W' = 1$	$W' = 10$	$W' = 100$	$W' = 1000$
3	$0.0021s$	$0.0021s$	$0.0021s$	$0.0023s$
5	$0.0280s$	$0.0281s$	$0.0281s$	$0.0283s$
8	$1.6613s$	$1.6566s$	$1.6637s$	$1.6717s$

Table 3: EUPAC calculation time in seconds for different multinomial sampling sizes N_{MN} and seen regret windows W'

value of around 500, EUPAC raises up until around 20%. At around 20% of learning episodes, regrets start to get worse from around 400 to around 1000. For those EUPAC immediately forms a local maximum. Now, although the following regret values get better again, EUPAC falls back and stays at 0%. Only after having seen more safe regrets across more learning episodes, EUPAC starts to rise again around 40% of learning episodes. EUPAC reacts immediately to worsening regrets, whereas several better regrets are needed to get EUPAC better again. The same arguments can be made between around 45% and 50%, and 60% and 70% of learning episodes.

Table 3 shows average calculation times for EUPAC by Interval Checking with differing N_{MN} and W' across 10 trials each. It is clear that the influence of N_{MN} dominates the little to no effect from changing W'. Note that with one EUPAC setting N_{MN} and W are fixed across evaluations. This leads EUPAC by Interval Checking to have a linear worst time complexity with $O(W')$ mainly due to binning.

Although these results are based on a preliminary simulated safety benchmark with pendulum systems, these properties should generalize well since the present regret function contains sufficient information about system safety. To validate this claim, future work should focus on applying EUPAC to other benchmarks. Since EUPAC is a measure that evaluates safety without interfering with the learning itself, one exciting direction of following works is the use of EUPAC in-the-loop.

References

[1] Javier García, Fern, and o Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

[2] Motoya Ohnishi, Li Wang, Gennaro Notomista, and Magnus Egerstedt. Barrier-certified adaptive reinforcement learning with applications to brushbot navigation. *IEEE Transactions on Robotics*, 35(5):1186–1205, October 2019.

[3] Jason Choi, Fernando Castañeda, Claire J. Tomlin, and Koushil Sreenath. Reinforcement learning for safety-critical control under model uncertainty, using control lyapunov functions and control barrier functions, 2020.

[4] Andrea Bajcsy, Somil Bansal, Eli Bronstein, Varun Tolani, and Claire J. Tomlin. An efficient reachability-based framework for provably safe autonomous navigation in unknown environments. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, December 2019.

[5] Thomas M. Schnieders and T. Richard. Ranking importance of exoskeleton design aspects. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1):1331–1335, September 2018.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

[6] Clement Gehring and Doina Precup. Smart exploration in reinforcement learning using absolute temporal difference errors. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '13, page 1037–1044. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[7] Rogier Koppejan and Shimon Whiteson. Neuroevolutionary reinforcement learning for generalized control of simulated helicopters. *Evolutionary Intelligence*, 4(4):219–241, October 2011.

[8] Javier Garcıa, Daniel Acera, and Fernando Fernández. Safe reinforcement learning through probabilistic policy reuse. *RLDM 2013*, page 14, 2013.

[9] John Bragg and Ibrahim Habli. What is acceptably safe for reinforcement learning? In *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37*, pages 418–430. Springer, 2018.

[10] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[11] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low bellman rank are pac-learnable. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 06–11 Aug 2017.

[12] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1638–1646, Bejing, China, 22–24 Jun 2014. PMLR.

[13] David A Winter. *Biomechanics and Motor Control of Human Movement*. John Wiley & Sons, 2009.

[14] Mingxing Lyu, Weihai Chen, Xilun Ding, Jianhua Wang, Shaoping Bai, and Huichao Ren. Design of a biologically inspired lower limb exoskeleton for human gait rehabilitation. *Review of Scientific Instruments*, 87(10):104301, October 2016.

[15] Stanley Plagenhoef, F Gaynor Evans, and Thomas Abdelnour. Anatomical data for analyzing human motion. *Research quarterly for exercise and sport*, 54(2):169–178, 1983.

[16] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, 2015.

[17] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897, Lille, France, 07–09 Jul 2015. PMLR.

[18] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018.

[19] Richard Cheng, Gábor Orosz, Richard M. Murray, and Joel W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3387–3395, July 2019.

[20] Ashley Hill, Antonin Raffin, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, Rene Traore, Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, and Yuhuai Wu. Stable baselines, 2018.