Never Worse, Mostly Better: Stable Policy Improvement in Deep Reinforcement Learning

Extended Abstract

Pranav Khanna
Indian Institute of Technology
India
pranikhanna1998@gmail.com

Guy Tennenholtz
Technion
Israel

Nadav Merlis
Technion
Israel

Shie Mannor
Technion, NVIDIA
Israel

Chen Tessler
NVIDIA
Israel

KEYWORDS

Reinforcement Learning; Deep Learning; Stability

ACM Reference Format:
Pranav Khanna, Guy Tennenholtz, Nadav Merlis, Shie Mannor, and Chen Tessler. 2023. Never Worse, Mostly Better: Stable Policy Improvement in Deep Reinforcement Learning: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

We **seek stable and always improving agents**. We aim to measure the reliability of each individual training run and suggest both *internal* and *external* stabilization methods to resolve these issues. We define internal stability as the intra-training behavior, i.e., stability of the learning process itself (the agent with respect to itself and its own historical behavior). On the other hand, external stability is measured with respect to an external benchmark policy, such as a human operator, a heuristic algorithm, or even a previously trained agent.

Internal stability is concerned with the agent and its historical behavior. Ideally, the agent should be monotonously improving [2, 10]. Nevertheless, the learning process of deep RL agents is characterized by frequent instabilities in performance. This is often unnoticed, as smoothed learning curves give an illusion of stability. Although the general trend is often improving on average, halting the agent at a random point may result in arbitrarily poor performance. Alternatively, a learner is often able to access an external benchmark policy. Instead of attempting to imitate this policy, one may utilize it as a stabilizing benchmark, requiring the agent always to perform better.

2 EVEREST

"nEVER woRsE moStly beTter" (EVEREST)[1], an off-policy method, alternates between learning and evaluation phases. During training, the agent collects data using an online policy and is periodically paused for evaluation. Internal stability is ensured through high-probability updates of a target network, and external stability through high-probability constraints of an admissible action set.

Internal Stability. We say that an RL algorithm is δ-internally-stable if, with probability at least $1 - \delta$, for every episode $k \in \{1, 2, \ldots\}$, $J^{\pi_k} \geq J^{\pi_{k-1}}$. A common approach in off-policy learning is to use target networks to stabilize learning [3, 4, 7], by slowly following the policy performance, often using Polyak-Rupert averaging [8]. We propose to condition this update according to the likelihood of improvement. Particularly, at every fixed interval, the learner proposes a new target network, which is updated if it improves upon the current target network with high probability.

External Stability. We say that an RL algorithm is δ-externally-stable w.r.t. a benchmark π_{ext} if, with probability at least $1 - \delta$, for every episode $k \in \{1, 2, \ldots\}$, $J^{\pi_k} \geq J^{\pi_{\text{ext}}}$. To achieve external stability, benchmark policies can be used to improve RL agents. The following theorem states that given a set of policies $\left\{\pi^{(i)}\right\}_{i=1}^{M}$, a per-state value maximizing policy, i.e., at each step playing the policy with the highest value, denoted by $\bar{\pi}$, achieves higher performance than any individual policy in the set.

Theorem 1. *Let* $\left\{\pi^{(i)}\right\}_{i=1}^{M}$ *and define* $\bar{\pi}$ *such that for all* $s \in \mathcal{S}$, $\bar{\pi}(s) \in \arg\max_{i\in[M]} v^{\pi^{(i)}}(s)$.

Then, $v^{\bar{\pi}}(s) \geq \max_{i\in[M]} v^{\pi^{(i)}}(s), \forall s \in \mathcal{S}$.

We utilize the result of Theorem 1 for the case of two policies – the learner π and the benchmark policy π_{ext}. To construct the mixture policy $\bar{\pi}$, we utilize an action elimination scheme; namely, at each state, the learner determines and constrains itself to the set of actions that improve its performance with high probability.

3 EXPERIMENTS

In this section we analyze our approach by focusing on three questions: (1) Do contemporary methods suffer from instability? (2) Does EVEREST empirically improve stability of these methods? (3) Does EVEREST improve the performance of these methods? In what follows we answer all three of these questions affirmatively.

Internal Stability: We compare the **(1) Baseline** TD3 algorithm with 3 variants of EVEREST (**(2) Max without reevaluation**, **(3) Max with reevaluation** and **(4) EVEREST**). While the baseline updates the target network at each step, using Polyak averaging,

[1] A full version of the paper is available at arxiv.org/abs/1910.01062

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

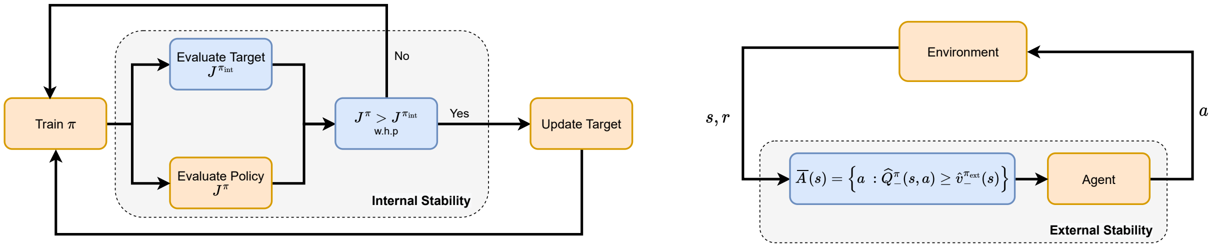
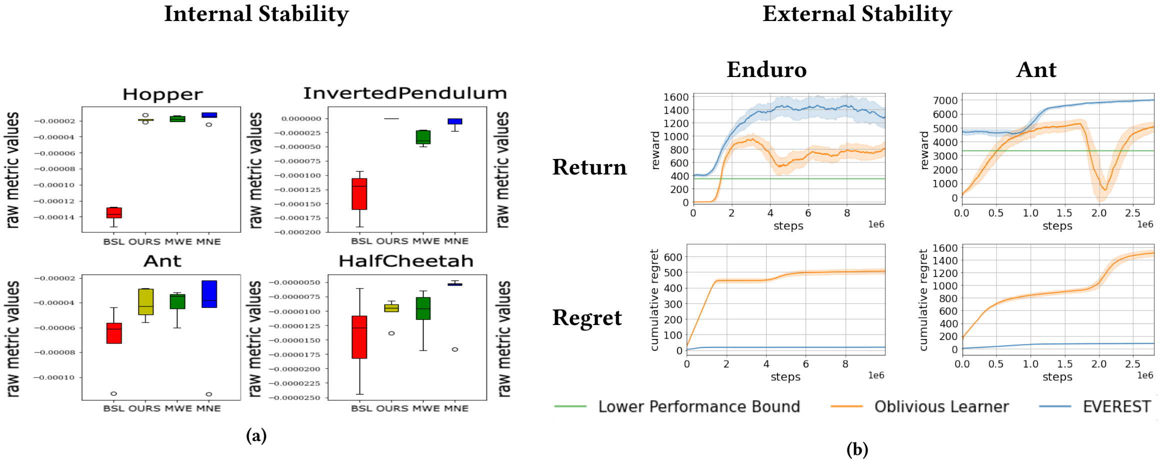

Train π
Evaluate Target $J^{\pi_{\text{int}}}$
Evaluate Policy J^{π}
$J^{\pi} > J^{\pi_{\text{int}}}$ w.h.p
No
Yes
Update Target
Internal Stability
Environment
s, r
a
$\overline{A}(s) = \left\{ a : \widehat{Q}^{\pi}_{-}(s,a) \geq \hat{v}^{\pi_{\text{ext}}}_{-}(s) \right\}$
Agent
External Stability

Figure 1: EVEREST assures internal stability (left) and external stability (right).



(a)

(b)

Figure 2: Internal stability: Reliability metrics for BSL (Oblivious TD3 Baseline), Ours (EVEREST), MWE (Max w/ reevaluation), and MNE (Max w/o reevaluation). We present the maximal average draw-down of the performance during training, measured as the CVaR over the 25th percentile. Better reliability is indicated by more positive values. External stability:

in (2) we update the target network based on the best performing historical network, in (3) we do so but re-evaluate the historical network to ensure an unbiased estimator, and in (4) we perform a proper statistical test to ensure high confidence improvement.

We measure the CVaR over de-trended differences. This shows the maximal expected drawdown (instability) of each method and the policies it proposes throughout training. Based on these results, all three variants significantly increase reliability compared with a baseline – showing the need to address internal stability and the benefit of changing the update scheme.

External Stability: An agent must guarantee that it will improve in accordance with a given benchmark policy with a high probability. In our experiments, we focus on pre-trained, sub-optimal agents. This is a common scenario in an evolving field such as RL, where the emergence of new techniques and improved compute result in better-performing policies. To illustrate such a scenario, our benchmark policies are obtained via partial training of a standard agent (SAC or DQN for MuJoCo and Atari, respectively). To stabilize the training process, EVEREST defines an admissible action set as the actions that improve upon the baseline with high probability.

Regret: As the admissible action set only contains actions that are w.h.p. at least as good as the benchmark, EVEREST exhibits lower regret across all tested scenarios. However, as EVEREST follows a probabilistic mechanism, it does not ensure zero violations. **Process performance:** An overly pessimistic agent may continually pass control to the benchmark and never become confident enough to take control and improve. What we observe is the opposite. Not only does EVEREST slowly take control and outperform the benchmark, but in all tasks, it exhibits performance at least as good as the oblivious learner. In addition, in some domains (Enduro and Ant) EVEREST exhibits superior performance compared to the oblivious learner (70% and 40% increase in performance, respectively).

4 CONCLUSIONS

RL is notoriously unstable and unreliable [5, 6]. As recent advances have focused on reproducibility, many of these concerns have been alleviated [1, 3, 4, 9]; however, since they emphasize the learning trend, they present smoothed learning curves, creating a false impression of stability.

EVEREST is a simple and theoretically justified method. It is shown to improve internal stability, producing more reliable and stable results, and when provided access to a baseline policy, improves external stability often outperforming a clean-slate agent.

REFERENCES

[1] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. 2020. Agent57: Outperforming the atari human benchmark. In *International Conference on Machine Learning*. PMLR, 507–517.

[2] Olivier J Bousquet, Amit Daniely, Haim Kaplan, Yishay Mansour, Shay Moran, and Uri Stemmer. 2022. Monotone Learning. In *Conference on Learning Theory*. PMLR, 842–866.

[3] Scott Fujimoto, Herke van Hoof, and David Meger. 2018. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477* (2018).

[4] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning*. 1856–1865.

[5] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[6] Alex Irpan. 2018. Deep Reinforcement Learning Doesn't Work Yet. https://www.alexirpan.com/2018/02/14/rl-hard.html.

[7] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.

[8] Boris T Polyak. 1990. New stochastic approximation type procedures. *Automat. i Telemekh* 7, 98-107 (1990), 2.

[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[10] Tom Viering, Alexander Mey, and Marco Loog. 2019. Open problem: Monotonicity of learning. In *Conference on Learning Theory*. PMLR, 3198–3201.