SAFE POLICY IMPROVEMENT WITH SOFT BASELINE BOOTSTRAPPING

Kimia Nadjahi*, Romain Laroche*, Rémi Tachet des Combes

kimia.nadjahi@telecom-paris.fr, {romain.laroche, remi.tachet}@microsoft.com

ABSTRACT

Problem setting

- Batch setting: fixed set of trajectories.
- Access to the behavioral policy called *baseline*.
- Objective: Improve the baseline with high probability.

Reliability issues in Reinforcement Learning

Deep Reinforcement Learning is unreliable.

- No convergence/optimality proof,
  → benchmark performance is the standard evaluation.
- Many, poorly understood hyperparameters,
  → not reproducible from one task to another.
- Very unstable learning process [1],
  → different seeds may yield very different results.

Batch Reinforcement Learning [2] is unreliable.

- Classic algorithms trained on a fixed dataset consider it as the environment (either explicitly or implicitly).
- This may be statistically insufficient, due to environmental stochasticity or function approximation.
- This leads to overestimates in the trained values.
- Classic algorithms perform planning, which turns out to be over-reliant on the overestimated states.

Safe Policy Improvement with Baseline Bootstrapping (SPIBB, [3])

- Recent computationally efficient and provably-safe methodology for batch RL.
- SPIBB updates the policy for frequent state-action pairs in the dataset only.
- SPIBB relies on a binary decision-making and may be too conservative.

Our contributions

- Novel batch RL method: *Soft-SPIBB*, a reformulation of SPIBB objective which makes it more flexible.
- Soft-SPIBB has proofs of safety guarantees and of computational efficiency in finite MDPs.
- Model-free Soft-SPIBB for function approximation.
- Empirical validation (performance, safety) on two domains.

THEORY

Safe Policy Improvement with Soft Baseline Bootstrapping (Soft-SPIBB)

- True environment $M^* = \langle \mathcal{X}, \mathcal{A}, P^*, R^*, \gamma \rangle$ is unknown,
- Maximum Likelihood Estimation (MLE) MDP built from counts: $\widehat{M} = \langle \mathcal{X}, \mathcal{A}, \widehat{P}, \widehat{R}, \gamma \rangle$.
- Error function e is derived from concentrations inequalities to bound the difference between parameters of $\widehat{M}$ and M^*: e_Q between Q-functions, e_P between probabilities.
- Soft-SPIBB relies on a softer mechanism where, for a given error function, a local error budget is allocated for policy changes in each state.
- A policy π is (π_b, e, ϵ)-constrained with hyper-parameter ϵ if, for each state $x \in \mathcal{X}$,
$$\sum_{a \in \mathcal{A}} e(x,a) \left| \pi(a|x) - \pi_b(a|x) \right| \leq \epsilon.$$
- A policy π is π_b-advantageous in $\widehat{M}$ if, for all $x \in \mathcal{X}$:
$$\sum_{a \in \mathcal{A}} \left(Q^{\pi_b}_{\widehat{M}}(x,a) - V^{\pi_b}_{\widehat{M}}(x) \right) \pi(a|x) \geq 0.$$
- Two algorithms performing policy iteration in the space of (π_b, e, ϵ)-constrained policies: *Exact-Soft-SPIBB* (exact solution) and *Approx-Soft-SPIBB* (tractable approximate solution).
- Model-free formulation, which fits the Q-function to the following targets:
$$y_j^{(i+1)} = r_j + \gamma \sum_{a' \in \mathcal{A}} \pi^{(i+1)}(a'|x'_j) Q^{(i)}(x'_j, a').$$

Theorems

Safe policy improvement bounds.

- Assume π is (π_b, e_Q, ϵ)-constrained and π_b-advantageous in $\widehat{M}$. For each state x, with high probability $1-\delta$:
$$V^{\pi}_{M^*}(x) - V^{\pi_b}_{M^*}(x) \geq -\frac{\epsilon V_{max}}{1-\gamma}.$$
- Assume π is (π_b, e_P, ϵ)-constrained. Suppose there exists $\kappa < \frac{1}{\gamma}$ such that,
$$\forall\, (x,a) \in \mathcal{X} \times \mathcal{A},\ \sum_{x',a'} e_P(x',a') \pi_b(a'|x') P^*(x'|x,a) \leq \kappa e_P(x,a).$$
We denote by $d_M(\cdot|x)$ the γ-discounted future state distribution starting from x when following π_b in M. Then, for each state x, with high probability $1-\delta$:
$$V^{\pi}_{M^*}(x) - V^{\pi_b}_{M^*}(x) \geq V^{\pi}_{\widehat{M}}(x) - V^{\pi_b}_{\widehat{M}}(x) - 2 \left\| d_{M^*}(\cdot|x) - d_{\widehat{M}}(\cdot|x) \right\|_1 V_{max} - \frac{1+\gamma}{(1-\gamma)^2 (1-\kappa\gamma)}\, \epsilon V_{max}.$$

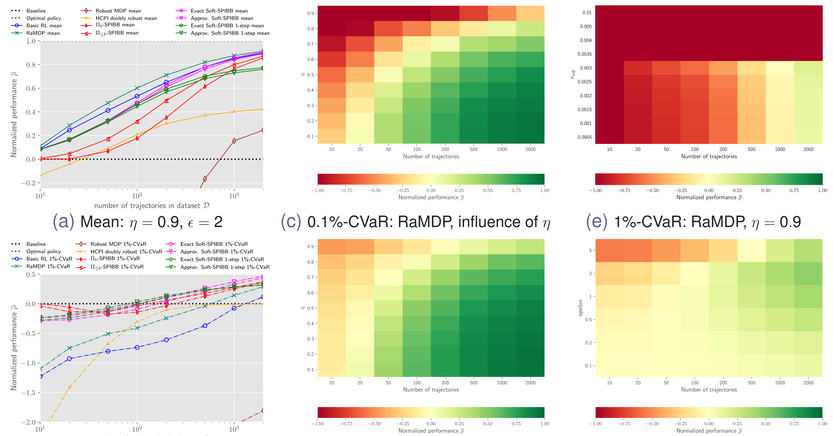
Analysis of Approx-Soft-SPIBB. The policy improvement step of Approx-Soft-SPIBB generates (π_b, e, ϵ)-constrained policies, and has a complexity of $\mathcal{O}(|\mathcal{X}||\mathcal{A}|^2)$.

Model-free formulation equivalence. In finite MDPs, the model-free policy iteration of (Exact or Approx)-Soft-SPIBB coincides with the model-based counterparts.

EXPERIMENTS

Random MDPs and baselines

Randomly generated finite MDPs and baselines (using a predefined level of performance η) → Normalized performance: $\overline{\rho}(\pi, M^*) = \dfrac{\rho(\pi, M^*) - \rho(\pi_b, M^*)}{\rho(\pi^*, M^*) - \rho(\pi_b, M^*)}$.

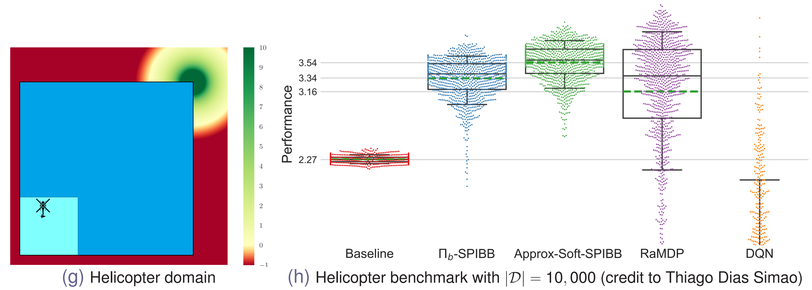


(a) Mean: $\eta = 0.9$, $\epsilon = 2$ (c) 0.1%-CVaR: RaMDP, influence of η (e) 1%-CVaR: RaMDP, $\eta = 0.9$

(b) 1%-CVaR: $\eta = 0.9$, $\epsilon = 2$ (d) 0.1%-CVaR: Approx-Soft-SPIBB, $\epsilon = 2$, influence of η (f) 1%-CVaR: Approx-Soft-SPIBB, $\eta = 0.9$

Figures (a-b) show the mean and CVaR benchmark. Figures (c-d) (resp. (e-f)) study the sensitivity to η (resp. to hyperparameters) for RaMDP [4] and Soft-SPIBB.

Helicopter domain (continuous task)



(g) Helicopter domain (h) Helicopter benchmark with $|\mathcal{D}| = 10,000$ (credit to Thiago Dias Simao)

REFERENCES

[1] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep RL that matters. In *AAAI*, 2018.
[2] S. Lange, T. Gabel, and M. Riedmiller. Batch Reinforcement Learning. In *Reinforcement Learning*. 2012.
[3] R. Laroche, P. Trichelair, and R. Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *ICML*, 2019.
[4] M. Petrik, M. Ghavamzadeh, and Y. Chow. Safe policy improvement by minimizing robust baseline regret. In *NIPS*, 2016.