Safe Policy Improvement with an Estimated Baseline Policy

Thiago D. Simão,¹ Romain Laroche,² Rémi Tachet des Combes²
¹Delft University of Technology, The Netherlands
²Microsoft Research Montréal

t.diasmiao@tudelft.nl, {romain.laroche, remi.tachet}@microsoft.com

Abstract

Previous work has shown the unreliability of existing algorithms in the batch Reinforcement Learning setting, and proposed the theoretically-grounded Safe Policy Improvement with Baseline Bootstrapping (SPIBB) fix: reproduce the baseline policy in the uncertain state-action pairs, in order to control the variance on the trained policy performance. However, in many real-world applications such as dialogue systems, pharmaceutical tests or crop management, data is collected under human supervision and the baseline remains unknown. In this paper, we apply SPIBB algorithms with a baseline estimate built from the data. We formally show safe policy improvement guarantees over the true baseline even without direct access to it. Our empirical experiments on finite and continuous states tasks support the theoretical findings. It shows little loss of performance in comparison with SPIBB when the baseline policy is given, and more importantly, drastically and significantly outperforms competing algorithms both in safe policy improvement, and in average performance.

1 Introduction

Reinforcement Learning (RL) is a framework for sequential decision-making optimization. Most RL research focuses on the online setting, where the system directly interacts with the environment and learns from it (Mnih et al. 2015; Van Seijen et al. 2017). While this setting might be the most efficient in simulation and in uni-device system control such as drones or complex industrial flow optimization, most real-world tasks (RWTs) involve a distributed architecture. We may cite a few: distributed devices (Internet of Things), mobile/computer applications (games, dialogue systems), or distributed lab experiments (pharmaceutical tests, crop management). These RWTs entail a high parallelization of the trajectory collection and strict communication constraints both in bandwidth and in privacy (Féraud, Alami, and Laroche 2019). Rather than spending a small amount of computational resource after each sample/trajectory collection, it is therefore more practical to collect a dataset using a behavioral (or baseline) policy, and then train a new policy from it. This setting is called batch RL (Lange, Gabel, and Riedmiller 2012).

Classically, batch RL algorithms apply dynamic programming on the samples in the dataset (Lagoudakis and Parr 2003; Ernst, Geurts, and Wènèkèl 2005). Laroche, Trichelair, and Tachet des Combes (2019) showed that in finite-state Markov Decision Processes (MDPs), these algorithms all converge to the same policy: the one that is optimal in the MDP with the maximum likelihood given the batch of data. Petrik, Ghavamzadeh, and Chow (2016) show that this policy is approximately optimal to the order of the inverse square root of the minimal state-action pairs count in the dataset. Unfortunately, Laroche, Trichelair, and Tachet des Combes (2019) show that even on very small tasks this minimal amount is almost always zero, and that, as a consequence, it gravely impairs the reliability of the approach: dynamic programming on the batch happens to return policies that perform terribly in the real environment. If a bad policy were to be run in distributed architectures such as the aforementioned ones, the consequences would be disastrous as it would jeopardize a high number of systems, or even lives.

Several attempts have been made to design reliable batch RL algorithms, starting with robust MDPs (Iyengar 2005; Nilim and El Ghaoû 2005), which consist of considering the set of plausible MDPs given the dataset, and then find the policy for which the minimal performance over the robust MDPs set is maximal. The algorithm however tends to converge to policies that are unnecessarily conservative.

Xu and Mannor (2009) considered robust regret over the optimal policy: the algorithm searches for the policy that minimizes the maximal gap with respect to the optimal performance in every MDP in the robust MDPs. However, they proved that evaluating the robust optimal regret for a fixed policy is already NP-complete with respect to the state and action sets’ size and the uncertainty constraints in the robust MDPs set.

Later, Petrik, Ghavamzadeh, and Chow (2016) considered the regret with respect to the behavioural policy performance over the robust MDPs set. The behavioural policy is called baseline in this context. Similarly, they proved that simply evaluating the robust baseline regret is already NP-complete. Concurrently, they also proposed, without theoretical grounding, the Reward-adjusted MDP algorithm (RaMDP), where the immediate reward for each transition in the batch is penalized by the inverse square root of the
number of samples in the dataset that have the same state and action than the considered transition.

Recently, Laroche, Trichet, and Tachet des Combes (2019) proposed Safe Policy Improvement with Baseline Bootstrapping (SPIBB), the first tractable algorithm with approximate policy improvement guarantees. Its principle consists in guaranteeing safe policy improvement by constraining the trained policy as follows: it has to reproduce the baseline policy in the uncertain state-action pairs. Nadjahi, Laroche, and Tachet des Combes (2019) further improved SPIBB’s empirical performance by adopting soft constraints instead. Related to this track of research, Simão Jahi, Laroche, and Tachet des Combes (2019) also developed SPIBB algorithms specifically for factored MDPs.

Concurrently to robust approaches described above, another tractable and theoretically-grounded family of frequentist algorithms appeared under the name of High Confidence Policy Improvement (Paduraru 2013; Mandel et al. 2014; Thomas, Theocharous, and Ghavamzadeh 2013a, b; HCPI), relying on importance sampling estimates of the trained policy performance. The algorithm in Mandel et al. (2014), based on concentration inequalities, tends to be conservative and requires hyper parameters optimization. The algorithms in Thomas, Theocharous, and Ghavamzadeh (2015b) rely on the assumption that the importance sampling estimate is normally distributed which is false when the number of trajectories is small. The algorithm in Paduraru (2013) is based on bias corrected and accelerated bootstrap and tends to be too optimistic. In contrast with the robust approaches, from robust MDPs to Soft-SPIBB, HCPI may be readily applied to infinite MDPs with guarantees. However, it is well known that the importance sampling estimates have high variance, exponential with the horizon of the MDP. The SPIBB algorithm has a linear horizon dependency, given a fixed known maximal value and the common horizon/discount factor equivalence: $H = \frac{1}{1 - \gamma}$ (Kocsis and Szepesvári 2006). Soft-SPIBB suffers a cubic upper bound but the empirical results rather indicate a linear dependency.

Nadjahi, Laroche, and Tachet des Combes (2019) perform a benchmark on randomly generated finite MDPs, baselines, and datasets. They report that the SPIBB and Soft-SPIBB algorithms are significantly the most reliable, and tie with RaMDP as the highest average performing algorithms. Additionally, they perform a benchmark on a continuous state space task, where the SPIBB and Soft-SPIBB algorithms significantly outperform RaMDP and Double-DQN (Van Hasselt, Guez, and Silver 2016) both in reliability and average performance. Soft-SPIBB particularly shines in the continuous state experiments.

Despite these appealing results, there is a caveat: the SPIBB and Soft-SPIBB algorithms rely on the knowledge of the baseline policy (like other robust baseline regret methods and HCPI). Although it may be available in some cases, it often happens that the dataset was collected using human interactions. For instance, research on dialogue systems strives after training reinforcement learning dialogue systems from human-human dialogues (Serban et al. 2016) while in medical data, a doctor’s prescriptions might be available, but not the policy that was followed. To overcome this issue, we investigate the use of SPIBB and Soft-SPIBB algorithms in the setting where the baseline policy is unknown.

Our contributions are threefold:

1. We formally prove safety bounds for SPIBB and Soft-SPIBB algorithms with estimated baseline policies in finite MDPs (Section 5).
2. We consolidate the theoretical results with empirical results in infinite randomly generated MDPs, unknown baselines, and datasets (Section 4).
3. We apply the method on a continuous state task by investigating two types of behavioural cloning, and show that it outperforms competing algorithms by a large margin, in particular on small datasets (Section 4).

Section 5 concludes the paper. A supplementary material is available to the interested reader with a glossary of our notations (Section A), links to anonymous github projects with the code of our experiments (Section B), additional empirical results/analysis for the finite MDPs (Section C) and for the continuous state task (Section D). But first, Section 2 recalls the necessary technical background.

## 2 Background

This section reviews the previous technical results relevant for this work.

### 2.1 Preliminaries

A Markov Decision Process (MDP) is the standard formalism to model sequential decision making problems in stochastic environments. An MDP $M$ is defined as $M = (\mathcal{X}, A, P, R, \gamma)$, where $\mathcal{X}$ is the state space, $A$ is the set of actions the agent can execute, $P: \mathcal{X} \times A \rightarrow \Delta_\mathcal{X}$ is the stochastic transition function, $R: \mathcal{X} \times A \rightarrow [-R_{\text{max}}, R_{\text{max}}]$, is a stochastic immediate reward function, $\gamma$ is the discount factor. Without loss of generality, we assume that the initial state is deterministic $x_t$.

A policy $\pi: \mathcal{X} \rightarrow \Delta_A$ represents how the agent interacts with the environment. The value of a policy $\pi$ starting from a state $x \in \mathcal{X}$ is given by the expected sum of discounted future rewards:

$$V^\pi_M(x) = \mathbb{E}_{x, M, x_0=x} \left[ \sum_{t \geq 0} \gamma^t R(x_t, a_t) \right]. \quad (1)$$

Therefore, the performance of a policy is the value in the initial state $x_t$. The goal of a reinforcement learning agent is to find a policy $\pi: \mathcal{X} \rightarrow \Delta_A$ that maximizes its expected sum of discounted rewards, however the agent does not have access to the dynamics of the true environment $M^* = (\mathcal{X}, A, P^*, R^*, \gamma)$.

In the batch RL setting, the algorithm receives as an input the dataset of previous transitions collected by executing a baseline policy $\pi_b: \mathcal{D} = \{x_k, a_k, r_k, x'_k, t_k\}_{k \in [1, |\mathcal{D}|]}$, where $x_k = x_i$ if $t_k = 0$ and $x_k = x_{k-1}$ otherwise is the starting state of the transition, $a_k \sim \pi_b(\cdot|x_k)$ is the performed action, $r_k \sim R(x_k, a_k)$ is the immediate reward, $x'_k \sim P(\cdot|x_k, a_k)$ is the reached state, and $t_k = 0$ if the previous transition was final and $t_k = t_{k-1} + 1$ otherwise is the trajectory-wise timestep.
We build from a dataset $\mathcal{D}$ the Maximum Likelihood Estimate (MLE) MDP $\hat{M} = (\mathcal{X}, \mathcal{A}, \hat{P}, \hat{R}, \gamma)$, as follows:

$$\hat{P}(x'|x, a) = \frac{N_D(x, a, x')}{N_D(x, a)},$$

$$\hat{R}(x, a) = \frac{\sum_{(x_j, a_j, x_j') \in \mathcal{D}} R_j}{N_D(x, a)},$$

where $N_D(x, a)$ is the state-action pair counts in the dataset $\mathcal{D}$. We also consider the robust MDPs set $\Xi$, i.e. the set of plausible MDPs such that the true environment MDP $M^*$ belongs to it with high probability $1 - \delta$:

$$\Xi = \{ M = (\mathcal{X}, \mathcal{A}, R, P, \gamma) \text{ s.t. } \forall x, a, (2) \}$$

where $e_b(x, a)$ is a model error function on the estimates of $\hat{M}$ for a state-action pair $(x, a)$. The error function is classically upper bounded with concentration inequalities.

In the next section, we discuss an objective for these algorithms that aims to guarantee a safe policy improvement for the new policy.

### 2.2 Approximate Safe Policy Improvement

Laroche, Trichelair, and Tachet des Combes (2019) investigate the setting where the agent receives as input the dataset $\mathcal{D}$ and must compute a new policy $\pi$ that approximately improves with high probability the baseline. Formally, the safety criterion can be defined as:

$$\mathbb{P}(\rho(\pi, M^*) \geq \rho(\pi_b, M^*) - \zeta) \geq \delta,$$

where $\zeta$ is a hyper-parameter indicating the improvement approximation and $1 - \delta$ is the high confidence hyper-parameter. Petrik, Ghavamzadeh, and Chow (2016) demonstrate that the optimization of this objective is NP-hard. To make the problem tractable, Laroche, Trichelair, and Tachet des Combes (2019) end up considering an approximate solution by maximizing the policy in the MLE-MDP while constraining the policy to be approximately improving in the robust MDPs set $\Xi$.

Given a hyper-parameter $N_\lambda$, their algorithm $\Pi_b$-SPIBB constrains the policy search to the set $\Pi_b$ of policies that reproduce the baseline probabilities in the state-action pairs that are present less than $N_\lambda$ times in the dataset $\mathcal{D}$:

$$\Pi_b = \{ \pi \text{ s.t. } \forall M \in \Xi, \rho(\pi, M) \geq \rho(\pi_b, M) - \zeta \}$$

We now recall the safe policy improvement guaranteed by $\Pi_b$-SPIBB:

**Theorem 1** (Safe policy improvement with baseline bootstrapping). Let $\pi_b^*$ be the optimal policy constrained to $\Pi_b$ in the MLE-MDP. Then, $\pi_b^*$ is a $\zeta$-approximate safe policy improvement over the baseline $\pi_b$ with high probability $1 - \delta$, where:

$$\zeta = 4V_{\max}\left(\frac{2}{N_\lambda}\log\frac{2|\mathcal{X}||\mathcal{A}|^{2|\mathcal{X}|}}{\delta} - \rho(\pi_b, \hat{M}) + \rho(\pi_b, \hat{M})\right)$$

Another algorithm considered in our work is Soft-SPIBB. Soft-SPIBB constrains the policy search such that the cumulative state-local error never exceeds $\epsilon$, with $\epsilon$ a fixed hyperparameter. More formally, the policy constraint is expressed as follows:

$$\Pi_{\epsilon} = \left\{ \pi \text{ s.t. } \forall x, a \in \mathcal{A}, \left| \pi(a|x) - \pi_b(a|x) \right| \leq \epsilon \right\}$$

Under some assumptions, Nadjah, Laroche, and Tachet des Combes (2019) demonstrate a looser safe policy improvement bound. Nevertheless, the policy search is less constrained and their empirical evaluation reveals that Soft-SPIBB safely finds better policies than SPIBB.

### 3 Baseline Estimates

In this section, we consider that the true baseline is unknown and implement a baseline estimate in order for the SPIBB and Soft-SPIBB algorithms to still be applicable.

#### 3.1 Algorithm and analysis

We construct the Maximum Likelihood Estimate of the baseline $\hat{\pi}_b$ (MLE baseline) as follows:

$$\hat{\pi}_b(a|x) = \begin{cases} \frac{N_D(x, a)}{|\mathcal{A}|} & \text{if } N_D(x) > 0, \\ \frac{1}{|\mathcal{A}|} & \text{otherwise}, \end{cases}$$

where $N_D(x)$ is the number of samples starting from state $x$ in dataset $\mathcal{D}$. Using this MLE policy, we may prove approximate safe policy improvement:

**Theorem 2** (Safe policy improvement with a baseline estimate). Assume given an algorithm $\alpha$ relying on the baseline $\pi_b$ to train a $\zeta$-approximate safe policy improvement $\pi_b^*$ over $\pi_b$ with high probability $1 - \delta$. Then, $\alpha$ with an MLE baseline $\hat{\pi}_b$ allows to train a $\hat{\zeta}$-approximate safe policy improvement $\hat{\pi}_b^*$ over $\pi_b$ with high probability $1 - \delta$:

$$\hat{\delta} = \delta + 2\epsilon'$$

$$\hat{\zeta} = \zeta + 2R_{\max}\sqrt{\frac{3|\mathcal{X}||\mathcal{A}| + 4\log \frac{1}{\delta'}}{2N}},$$

where $N$ is the number of trajectories in the dataset and $1 - \delta'$ controls the uncertainty stemming from the baseline estimation.

**Proof.** We are ultimately interested in the performance improvement of $\hat{\pi}_b^*$ with respect to the true baseline $\pi_b$ in the true environment. To do so, we decompose the difference into two parts:

$$\rho(\pi_b^*, M^*) - \rho(\pi_b, M^*) = \rho(\hat{\pi}_b^*, M^*) - \rho(\pi_b, M^*) + \rho(\hat{\pi}_b, M^*) - \rho(\pi_b, M^*).$$

Regarding the first term, note that, while $\hat{\pi}_b$ is not the true baseline, it is the MLE baseline, meaning in particular that it was more likely to generate the dataset $\mathcal{D}$ than the true one.
Hence, we may consider it as a potential behaviour policy, and apply the safe policy improvement guarantee provided by algorithm \(\alpha\) to bound the difference.

Regarding the second term, we need to use the distributional formulation of the performance of any policy \(\pi\):

\[
\rho(\pi, M) = \sum_{x \in X} \sum_{a \in A} d_M^\pi(x, a) \mathbb{E} R(x, a),
\]

where \(d_M^\pi(x, a)\) is the discounted sum of visits of state-action pair \((x, a)\) while following policy \(\pi\) in MDP \(M\). Then, we may rewrite the second term in Equation 9 and upper bound it using Hölder’s inequality as follows:

\[
\sum_{x \in X} \sum_{a \in A} \left( d_M^{\pi^*}(x, a) - d_M^\pi(x, a) \right)^2 \mathbb{E} R^4(x, a) \\
\leq \left\| d_M^{\pi^*}(x, a) - d_M^\pi(x, a) \right\|_1 R_{\max}. \tag{11}
\]

Next, we decompose the state-action discounted visits divergence as follows:

\[
\left\| d_M^{\pi^*}(x, a) - d_M^\pi(x, a) \right\|_1 \leq \left\| d_M^{\pi^*}(x, a) - d_D(x, a) \right\|_1 + \left\| d_M^\pi(x, a) - d_D(x, a) \right\|_1,
\]

where \(d_D(x, a)\) is the state-action discounted visits in dataset \(D\). Regarding the first term, Weissman et al. (2003) provide the following concentration inequality:

\[
\mathbb{P} \left( \left\| d_M^{\pi^*}(x, a) - d_D(x, a) \right\|_1 (1 - \gamma) \geq \varepsilon \right) \\
\leq \left( 2|X||A| - 2 \right) \exp \left( -\frac{N\varepsilon^2}{2} \right), \tag{13}
\]

where \(N\) is the number of trajectories in the dataset, and \(\varepsilon\) is an intermediate variable we introduce. With a little calculus and by setting the right value to \(\varepsilon\), we then obtain with high probability \(1 - \delta^*\):

\[
\left\| d_M^{\pi^*}(x, a) - d_D(x, a) \right\|_1 \leq \frac{1}{1-\gamma} \sqrt{\frac{3|X||A| + 4 \log \frac{1}{\delta^*}}{2N}}.
\]

Regarding the second term of Equation 12, we may observe that there is a correlation between \(d_M^{\pi^*}\) and \(d_D\) through \(D\), but it is a positive correlation, meaning that the divergence between the distributions is smaller than the one with an independently drawn dataset of the same size. As a consequence, we are also able to upper bound it by assuming independence, and using the same development as for the first term. This finally gives us from Equation 12 and with high probability \(1 - 2\delta^*\):

\[
\left\| d_M^{\pi^*}(x, a) - d_M^\pi(x, a) \right\|_1 \leq \frac{2}{1-\gamma} \sqrt{\frac{3|X||A| + 4 \log \frac{1}{\delta^*}}{2N}}, \tag{14}
\]

which allows us to conclude the proof using union bounds.

\footnote{We need to rescale with \((1 - \gamma)\) the state-action discounted visits to make it sum to 1 since the original bound applies to probability distributions.}

### 3.2 Theorem 2 discussion

SPIBB and Soft-SPIBB safe policy improvement guarantees exhibit a trade-off (controlled with their respective hyper-parameters \(\frac{1}{\sqrt{M}}\) and \(\varepsilon\)) between upper bounding the true policy improvement error (first term in Theorem 1) and allowing maximal policy improvement in the MLE MDP (next terms). When the hyper-parameters are set to 0, the true policy improvement error is null, because, trivially, no policy improvement is allowed: the algorithm is forced to reproduce the baseline. When the hyper-parameters grow, larger improvements are permitted, but the error upper bound term also grows. When the hyper-parameters tend to \(+\infty\), the algorithms are not constrained anymore and find the optimal policy in the MLE MDP. In that case, the error is no longer upper bounded, resulting in poor safety performance.

When using the MLE baseline instead of the true baseline, Theorem 2 introduces another error upper bound term accounting for the accurateness of the baseline estimate that cannot be reduced by hyper-parameter settings. That fact is entirely expected, as otherwise we could consider an empty dataset, pretend it was generated with an optimal policy and expect a safe policy improvement over it. Another interesting point is that the bound depends on the number of trajectories, not the number of state-action visits, nor the total number of samples. Indeed, even with a huge number of samples, if there were collected only from a few trajectories, the variance may still be high, since future states visited on the trajectory depend on the previous transitions.

Regarding the MDP parameters dependency, the upper bound grows as the square root of the state set size, as for standard SPIBB, but also grows as the square root of the action set size contrarily to SPIBB with has a logarithmic dependency, which may cause issues in some RL problems. The direct horizon dependency is the same (linear). But one could argue that it is actually lower. The maximal value \(V_{\max}\) in the SPIBB bounds can reach \(\frac{H_{\max}}{1-\gamma}\), making the total dependency \(\approx H^{3/2}\). In both cases, those are better than the Soft-SPIBB cubic dependency.

One may consider other baseline estimates than the MLE, using Bayesian priors for instance, and infer new bounds. This should work as long as the baseline estimate remains a policy that could have generated the dataset.

### 4 Empirical Analysis

We split our empirical analysis in two parts, the first considers random MDPs with finite state spaces and the second MDPs with continuous state spaces.

#### 4.1 Random finite MDPs

**Setup:** This experiment has for objective to empirically analyse the consistency between the theoretical findings and the practice. To do so, and in order to account to a broad range of tasks, we reproduce the experiment found in [Laroche, Trichet, and Tachet des Combes](2019) and
Finite MDPs with \( \eta = 0.9, N_A = 7 \) and \( \epsilon = 0.5 \). On the left, the mean curves, on the right, the 1%-quantile curves.

Nadjahi, Laroche, and Tachet des Combes (2019). The experiment is run on finite MDPs that are randomly generated, with randomly generated baseline policies from which trajectories are obtained. We quickly recall the setting below.

The true environment is a randomly generated MDP with 50 states, 4 actions, and a transition connectivity of 4: a given state-action pair may transit to four different states at most. The reward function is 0 everywhere except for the transitions entering the target state, in which case the trajectory terminates with a reward of 1. The target state is the hardest state to reach from the initial one.

The baselines are also randomly generated with a predefined level of performance specified by a ratio \( \eta \) between the optimal policy \( \pi^* \) performance and the uniform policy \( \hat{\pi} \) performance: \( \rho(\pi_b, M) = \eta \rho(\pi^*, M) + (1 - \eta) \rho(\hat{\pi}, M) \). For more details on the process, we refer the interested reader to the original papers. Two values for \( \eta \) are considered: the experiments with \( \eta = 0.9 \) are reported in the main document, and the experiments with \( \eta = 0.1 \) are reported in the supplementary material, Section C. We also study the influence of the dataset size \(|D|\in[10,20,50,100,200,500,1000,2000] \).

Competing algorithms: Our plots display nine curves:
- \( \pi^* \): the optimal policy,
- \( \pi_b \): the true baseline,
- \( \hat{\pi}_b \): the MLE baseline,
- \( \Pi_b\text{-SPIBB} \): SPIBB algorithm on the true baseline,
- \( \hat{\Pi}_b\text{-SPIBB} \): SPIBB algorithm on the MLE baseline,
- \( \hat{\Pi}_b\text{-Soft} \): Soft-SPIBB algorithm on the true baseline,
- \( \hat{\Pi}_b\text{-Soft} \): Soft-SPIBB algorithm on the MLE baseline,
- RaMDP: Reward-adjusted MDP,
- and Basic RL: dynamic programming on the MLE MDP.

All the algorithms are compared using their optimal hyper-parameter according to previous work. Our hyper-parameter search with the MLE baselines did not show significant differences and we opted to report results with the same hyper-parameter values. Soft-SPIBB algorithms are the ones coined as Approx. Soft SPIBB in Nadjahi, Laroche, and Tachet des Combes (2019).

Performance indicators: Given the random nature of the MDP and baseline generations, we need to normalize the performance to allow inter-experiment comparison:

\[
\rho = \frac{\rho(\pi, M^*) - \rho(\pi_b, M^*)}{\rho(\pi^*, M^*) - \rho(\pi_b, M^*)}.
\]

Thus, the optimal policy always has a normalized performance of 1, and the true baseline a normalized performance of 0. A positive normalized performance means a policy improvement, and a negative normalized performance an infringement of the policy improvement objective. Figures either report the average normalized performance of the algorithms or its 1%-quantile. Each setting is processed on 250k seeds, to ensure that every performance gap visible to the naked eye is significant.

Empirical results: Figure 1 shows the results with \( \eta = 0.9 \), i.e. the hard setting where the behavior baseline is almost optimal, and therefore difficult to improve.

Performance of the MLE baseline. First, we notice that the mean performance of the MLE baseline \( \hat{\pi}_b \) is slightly lower than the true baseline policy \( \pi_b \) for small datasets. As \(|D|\) increases, the performance of \( \hat{\pi}_b \) quickly increases to reach the same level. The 1%-quantile is significantly lower when the number of trajectories is reduced.

Soft-SPIBB with true and estimated baselines. Comparing the results of \( \Pi_b\text{-Soft} \) and \( \hat{\Pi}_b\text{-Soft} \) curves, it is surprising that the policy computed using an estimated policy as a baseline yields better results than the one computed with the true policy. Hanna, Niekum, and Stone (2019) reported similar results in a different setting and formally explained them with the sampling error.

SPIBB with true and estimated baselines. Analysing the performance of the \( \hat{\Pi}_b\text{-SPIBB} \) algorithm, we notice that it also slightly improves over \( \Pi_b\text{-SPIBB} \) on the mean normalized performance. As far as safety is concerned, we see that

\( \text{Note the difference with previously reported results in SPIBB papers, which focused on the conditional value at risk indicator.} \)
Figure 2: $|\mathcal{D}| = 10,000$. The green dashed line shows the average and the caps show the 10% and 90% percentile. Each dot on the swarm plots displays the evaluation of a seed.

the 1%-quantile of policies computed with $\hat{\pi}_b$-SPIBB falls close to the 1%-quantile of the estimated baseline $\bar{\pi}_b$ for small datasets and close to the 1%-quantile of the policies $\Pi_b$-SPIBB for datasets with around 100 trajectories. It is expected as $\hat{\pi}_b$-SPIBB tends to reproduce the baseline for very small datasets, and improves over it for larger ones. That statement is also true of $\hat{\pi}_b$-Soft.

RaMDP and Basic RL. Finally, it is interesting to observe that although RaMDP and Basic RL can compute policies with rather high mean performance, these algorithms often return policies performing much worse than the MLE policy $\bar{\pi}_b$ (as seen in their 1%-quantile).

4.2 Continuous MDPs

Helicopter domain: For MDPs with continuous state space, we focus on the helicopter environment [Laroche, Trichelair, and Tachet des Combes 2019, Figure 3]. In this stochastic domain, the state is defined by the position and velocity of the helicopter. The agent has a discrete set of 9 actions to control the thrust applied in each dimension. The helicopter begins in a random position of the bottom-left corner with a random initial velocity. The episode ends if the helicopter’s speed exceeds some threshold, giving a reward of -1, or if it leaves the valid region, in which case the agent gets a reward between -1 and 10 depending on how close it is to the top-right corner. Using a fixed behavior policy $\pi_b$, we generate 1,000 datasets for each algorithm. We report results for two dataset sizes: 3,000 and 10,000 transitions (approximately 3 times less trajectories).

Behavioural cloning: In infinite MDPs, there is no MLE baseline definition. We have to lean on behavioural cloning techniques. We compare here two straightforward ones in addition to the true behavior policy $\pi_b$: a baseline estimate $\bar{\pi}_c$ based on the pseudo-counts used by the algorithms, and a neural-based baseline estimate $\tilde{\pi}_n$ that uses a standard probabilistic classifier (see Section D of the supplementary material for more details).

Competing algorithms:
- $\pi_b$: the true baseline,
- $\bar{\pi}_c$: the pseudo-count-based estimate of the baseline,
- $\tilde{\pi}_n$: the neural-based estimate of the baseline,
- $\Pi_b$-SPIBB: SPIBB algorithm on the true baseline,
Table 1: Numerical results for the two size of datasets. The key performance indicators are respectively the percentage of policy improvement over the true baseline, the average performance of the trained policies, the 10%-quantile, and the 1%-quantile. For each column, we bold the best performing algorithm that is not using the true baseline $\pi_b$.

| Baseline | Algorithm | $|D| = 3,000$ | $|D| = 10,000$ |
|-------|-----------|--------------|--------------|
| $\pi_b$ |           | $\mathbb{P}(\rho(\pi) > \rho(\pi_b))$ | $\mathbb{P}(\rho(\pi) > \rho(\pi_b))$ |
|       |           | avg perf | 10%-quantile | 1%-quantile | avg perf | 10%-quantile | 1%-quantile |
| $\pi_n$ baseline | 0.499 | 2.27 | 2.22 | 2.18 | 0.499 | 2.27 | 2.22 | 2.18 |
| $\bar{\pi}_n$ Soft-SPIBB | 0.582 | 2.29 | 1.86 | 1.43 | 0.973 | 2.97 | 2.61 | 2.15 |
| $\bar{\pi}_c$ | 0.514 | 2.23 | 1.73 | 1.21 | 0.930 | 2.75 | 2.37 | 1.75 |
| $\pi_n$ Soft-SPIBB | 0.760 | 2.48 | 2.12 | 1.71 | 0.996 | 3.30 | 2.93 | 2.47 |
| $\bar{\pi}_c$ | 0.785 | 2.66 | 2.11 | 1.51 | 0.980 | 3.45 | 2.93 | 2.09 |
| N/A RaMDP | 0.006 | 0.37 | -0.75 | -0.99 | 0.876 | 3.16 | 2.13 | 0.23 |
| N/A Double-DQN | 0.001 | -0.77 | -1.00 | -1.00 | 0.076 | 0.25 | -0.97 | -1.00 |

Table 1: Numerical results for the two size of datasets. The key performance indicators are respectively the percentage of policy improvement over the true baseline, the average performance of the trained policies, the 10%-quantile, and the 1%-quantile. For each column, we bold the best performing algorithm that is not using the true baseline $\pi_b$.

- $\hat{\pi}_c$: SPIBB: SPIBB algorithm on $\hat{\pi}_c$.
- $\hat{\pi}_n$: SPIBB: SPIBB algorithm on $\hat{\pi}_n$.
- $\bar{\pi}_b$: Soft: Soft-SPIBB algorithm on the true baseline.
- $\bar{\pi}_c$: Soft: Soft-SPIBB algorithm on $\hat{\pi}_c$.
- $\bar{\pi}_n$: Soft: Soft-SPIBB algorithm on $\hat{\pi}_n$.
- RaMDP: Double-DQN with Reward-adjusted MDP.
- and Double-DQN: basic deep RL algorithm.

Performance indicators: The plots represent for each algorithm a modified box-plot where the caps show the 10%-quantile and 90%-quantile, the upper and lower limits of the box are the 25% and 75% quantiles and the middle line in black shows the median. We also show the average of each algorithm (dashed lines in green) and finally add a swarm-plot to enhance the distribution visualization.

The table provides additional details, including the percentage of policies that showed a performance above the average performance of the true baseline policy.

Results: The results are reported numerically in Table 1 and graphically on Figure 4 for $|D| = 10,000$ (for $|D| = 3,000$, refer to Section D of the supplementary material).

Empiric baseline policies. On Figure 2 we observe that the baseline policies $\pi_c$ and $\pi_n$ have a performance poorer than the true behavior policy $\pi_b$. On the one hand, the neural-based baseline estimate $\pi_n$ can get values close to the performance of the true behavior policy, however, it has a high variance and even the $90\%$-quantile is below the mean of the true policy. On the other hand, the count-based policy $\pi_c$ has a low variance, but it has a much lower mean performance. In general, we observe a much bigger performance loss than in finite MDPs between the true baseline and the estimated baseline.

SPIBB results. With SPIBB, the neural-based baseline estimate leads to better results for all indicators. The loss in average performance makes it worse than RaMDP in the $|D| = 10,000$ datasets, but it is more reliable and yields more consistently to policy improvements. On the $|D| = 3,000$ datasets, it demonstrates a higher robustness with respect to the small datasets, still compared to RaMDP.

Soft-SPIBB results. The Soft-SPIBB results with baseline estimates are impressive. The loss of performance with respect to Soft-SPIBB with the true baseline is minor. We highlight that, although the policy based on pseudo-counts has a lower performance than the true one (1 point difference), it still achieves a strong performance when used with Soft-SPIBB (less than 0.1 point difference). This indicates that the proposed method is robust with respect to the performance of the empirical policy. It seems that the soft policy change allowed by Soft-SPIBB makes a strong difference when the baseline is estimated.

5 Conclusion
This paper addresses the problem of performing safe policy improvement in batch RL without direct access to the baseline, i.e. the behavioural policy of the dataset. We provide the first theoretical guarantees for safe policy improvement in this setting, and show on finite and continuous MDPs that the algorithm is tractable and significantly outperforms all competing algorithms that do not have access to the baseline. We also empirically confirm the limits of the approach when the number of trajectories in the dataset is low.

Future work includes addressing the multi-batch setting, when there are several sequential updates (Laroche and Tachet des Combes 2019), extending the method to continuous action spaces, and investigating the use of SPIBB in a full online setting, as a value estimation stabilizer.
References

[Ernst, Geurts, and Wehenkel 2005] Ernst, D.; Geurts, P.; and Wehenkel, L. 2005. Tree-based batch mode reinforcement learning. Journal of Machine Learning Research 6(Apr):503–556.

[Féraud, Alami, and Laroche 2019] Féraud, R.; Alami, R.; and Laroche, R. 2019. Decentralized exploration in multi-armed bandits. In Proceedings of the 36th International Conference on Machine Learning (ICML, long oral).

[Hanna, Niekum, and Stone 2019] Hanna, J.; Niekum, S.; and Stone, P. 2019. Importance Sampling Policy Evaluation with an Estimated Behavior Policy. In Proceedings of the 36th International Conference on Machine Learning, ICML, 2605–2613. PMLR.

[Iyengar 2005] Iyengar, G. N. 2005. Robust dynamic programming. Mathematics of Operations Research.

[Kocsis and Szepesvári 2006] Kocsis, L., and Szepesvári, C. 2006. Bandit based monte-carlo planning. In Proceedings of the 4th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), 282–293. Springer.

[Lagoudakis and Parr 2003] Lagoudakis, M. G., and Parr, R. 2003. Least-squares policy iteration. Journal of machine learning research 4(Dec):1107–1149.

[Lange, Gabel, and Riedmiller 2012] Lange, S.; Gabel, T.; and Riedmiller, M. 2012. Batch Reinforcement Learning. Berlin, Heidelberg: Springer Berlin Heidelberg. 45–73.

[Laroche and Tachet des Combes 2019] Laroche, R., and Tachet des Combes, R. 2019. Multi-batch reinforcement learning. In Proceedings of the 4th Reinforcement Learning and Decision Making (RLDM).

[Laroche, Trichelair, and Tachet des Combes 2019] Laroche, R.; Trichelair, P.; and Tachet des Combes, R. 2019. Safe policy improvement with baseline bootstrapping. In Proceedings of the 36th International Conference on Machine Learning (ICML).

[Mandel et al. 2014] Mandel, T.; Liu, Y.-E.; Levine, S.; Brunskill, E.; and Popovic, Z. 2014. Offline policy evaluation across representations with applications to educational games. In Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS).

[Mnih et al. 2015] Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. Nature 518(7540):529.

[Nadjahi, Laroche, and Tachet des Combes 2019] Nadjahi, K.; Laroche, R.; and Tachet des Combes, R. 2019. Safe policy improvement with soft baseline bootstrapping. In Proceedings of the 17th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD).

[Nilim and El Ghaoui 2005] Nilim, A., and El Ghaoui, L. 2005. Robust control of markov decision processes with uncertain transition matrices. Operations Research.

[Paduraru 2013] Paduraru, C. 2013. Off-policy Evaluation in Markov Decision Processes. Ph.D. Dissertation, PhD thesis, McGill University.

[Petrik, Ghavamzadeh, and Chow 2016] Petrik, M.; Ghavamzadeh, M.; and Chow, Y. 2016. Safe policy improvement by minimizing robust baseline regret. In Proceedings of the 29th Advances in Neural Information Processing Systems (NIPS).

[Serban et al. 2016] Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In Thirtieth AAAI Conference on Artificial Intelligence.

[Simão and Spaan 2019a] Simão, T. D., and Spaan, M. T. J. 2019a. Safe policy improvement with baseline bootstrapping in factored environments. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence.

[Simão and Spaan 2019b] Simão, T. D., and Spaan, M. T. J. 2019b. Structure learning for safe policy improvement. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI).

[Thomas, Theocharous, and Ghavamzadeh 2015a] Thomas, P.; Theocharous, G.; and Ghavamzadeh, M. 2015a. High confidence policy improvement. In Proceedings of the 32nd International Conference on Machine Learning (ICML).

[Thomas, Theocharous, and Ghavamzadeh 2015b] Thomas, P. S.; Theocharous, G.; and Ghavamzadeh, M. 2015b. High-confidence off-policy evaluation. In Proceedings of the 29th AAAI Conference on Artificial Intelligence.

[Van Hasselt, Guez, and Silver 2016] Van Hasselt, H.; Guez, A.; and Silver, D. 2016. Deep reinforcement learning with double q-learning. In Proceedings of the 30th AAAI Conference on Artificial Intelligence.

[Van Seijen et al. 2017] Van Seijen, H.; Fatemi, M.; Laroche, R.; Romoff, J.; Barnes, T.; and Tsang, J. 2017. Hybrid reward architecture for reinforcement learning. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS).

[Weissman et al. 2003] Weissman, T.; Ordentlich, E.; Seroussi, G.; Verdu, S.; and Weinberger, M. J. 2003. Inequalities for the l1 deviation of the empirical distribution. Hewlett-Packard Labs, Tech. Rep.

[Xu and Mannor 2009] Xu, H., and Mannor, S. 2009. Parametric regret in uncertain markov decision processes. In Proceedings of the 48th IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference, 3606–3613. IEEE.
## Glossary

| Symbol       | Designation                                                                 |
|--------------|-------------------------------------------------------------------------------|
| $\mathbb{E}_z[Y]$ | Expectation of random variable $Y$ with variable $z$ fixed                    |
| $P(\square)$ | Probability that $\square$ is true                                           |
| $M$          | Markov Decision Process                                                       |
| $\mathcal{X}$ | State set                                                                    |
| $\mathcal{A}$ | Action set                                                                   |
| $P(x'|x,a)$   | Transition function                                                           |
| $R(x,a)$     | Reward function                                                               |
| $\gamma$     | Discount factor                                                               |
| $\hat{\square}$ | Optimal or true value of $\square$                                            |
| $\Delta_\square$ | Set of probability distributions over $\square$                             |
| $k$          | Sample index                                                                  |
| $x_i$        | Initial state                                                                 |
| $x_k$        | State of sample $k$                                                           |
| $a_k$        | Action taken in sample $k$                                                    |
| $x'_k$       | State accessed in sample $k$                                                  |
| $r_k$        | Immediate reward received in sample $k$                                       |
| $t_k$        | Trajectory wise time-step of sample $k$                                       |
| $R_{\text{max}}$ | Known bound of immediate rewards                                            |
| $V_{\text{max}}$ | Known bound of value                                                          |
| $\square$    | Size of finite set/list/collection $\square$                                  |
| $D$          | Dataset                                                                       |
| $\pi(a|x)$   | Policy                                                                        |
| $\hat{\pi}_b(a|x)$ | Count-based estimate of the baseline                                      |
| $\hat{\pi}_n(a|x)$ | Neural-based estimate of the baseline                                      |
| $\alpha$     | Algorithm                                                                     |
| $1 - \delta$ | High probability hyper-parameter                                               |
| $e_{\delta}(x,a)$ | Error bound of the model in state-action $(x,a)$ with high probability $1 - \delta$ |
| $\Xi$        | Robust MDPs set                                                               |
| $\zeta$      | Approximation of the safe policy improvement                                 |
| $N_{\wedge}$ | SPIBB hyper-parameter                                                         |
| $\epsilon$  | Soft-SPIBB hyper-parameter                                                    |
| $N$          | Trajectory count                                                               |
| $N_D(x)$     | State count                                                                   |
| $N_D(x,a)$   | State-action count                                                            |
| $N_D(x,a,x')$ | Transition count                                                              |
| $V^\pi_M(x)$ | Value of state $x$ in MDP $M$, while following policy $\pi$                   |
| $\rho(\pi,M)$ | Expected performance of policy $\pi$ in MDP $M$                               |
| $\rho$       | Normalized expected performance of policy $\pi$                              |
| $\eta$       | Hyper-parameter for the random baseline generation                            |
| $d^*_M(x,a)$ | Discounted state-action visits with policy $\pi$ in MDP $M$                   |
| $d_D(x,a)$   | Empirical discounted state-action visits                                       |

## Source Code

Our code is building upon the code published in previous SPIBB papers. The source code for the finite MDPs experiments is available at

https://github.com/paper4263/paper4263

The source code for the experiments with continuous state space MDPs is available at

https://github.com/paper4263/paper4263b
C  Random Finite MDPs

Figure 4 shows the results for Finite MDPs with the easy setting, where the performance of the baseline policy is low ($\eta = 0.1$) and it is easy to find an improved policy. We notice that all algorithms show a reliable performance improvement over the baseline. The estimated policy have a 1%-quantile slightly below the true behavior policy with small datasets and, naturally, the $\hat{\Pi}_b$-SPIBB shows a similar behavior.

D  MDPs with continuous state space

D.1 Baseline Estimate in MDPs with continuous state space

The count-based policy follows a principle similar to the MLE policy. It uses a pseudo-count for state-action pairs $\tilde{N}(x, a)$ defined according to the sum of the euclidean distance $\|x - x'\|_2$ from the state $x$ and all states of transitions in the dataset where the action $a$ was executed [Laroche, Trichet, and Tachet des Combes 2019 Section 3.4):

$$\tilde{N}_D(x, a) = \sum_{(x_j, a_j = a, r_j, x_j') \in \mathcal{D}} \max\{0, 1 - \frac{\|x - x_j\|_2}{d_0}\},$$  \hspace{1cm} (16)

where $d_0$ is a hyper-parameter to impose a minimum similarity before increasing the counter of a certain state. We also compute the state pseudo-count using this principle: $\tilde{N}_D(x) = \sum_{a \in A} \tilde{N}_D(x, a)$. This way, we can define the count-based baseline estimate replacing the count in Equation 6 by its pseudo-count counterpart:

$$\tilde{\pi}_c(a|x) = \frac{N_D(x, a)}{\tilde{N}_D(x)}, \text{ if } \tilde{N}_D(x) > 0, \hspace{1cm} \text{ (17)}$$

The neural-based estimate of the baseline $\tilde{\pi}_n(a|x)$ is estimated using a supervised learning approach. We train a probabilistic classifier using a neural network to minimize the negative log-likelihood with respect to the actions in the dataset.

We use the same architecture as the one used to train the Double-DQN models, which is shared among all the algorithms in the helicopter domain experiments: a fully connected neural network with 3 hidden layers of 32, 128 and 28 neurons respectively, and 9 outputs corresponding to the 9 actions.

To avoid overfitting, we split the data set in two parts, using 80% for training and 20% for validation. During training, we evaluate the classifier on the validation dataset at the end of every epoch and save the parameters of the network with the smallest validation loss.

D.2 Hyper-parameters

Building on the results presented by Nadja, Laroche, and Tachet des Combes (2019), we set the hyper-parameters for the experiments with $|\mathcal{D}| = 10,000$ ($|\mathcal{D}| = 3,000$) as follows:

- $\Pi_b$-SPIBB with $N_\lambda = 3$ ($N_\lambda = 1$),
- $\Pi_b$-Soft with $\epsilon = 0.6$ ($\epsilon = 0.8$),
- RaMPD with $\kappa = 1$ ($\kappa = 1.75$).
For the algorithms using an estimated baseline we run a parameter search (Section D.4) and set the parameters for the main experiments as follows:

- \( \hat{\Pi}_n \)-SPIBB with \( N_\lambda = 3.0 \) \( (N_\lambda = 1.0) \),
- \( \hat{\Pi}_c \)-SPIBB with \( N_\lambda = 3.0 \) \( (N_\lambda = 1.0) \),
- \( \hat{\Pi}_n \)-Soft with \( \epsilon = 0.6 \) \( (\epsilon = 0.8) \),
- \( \hat{\Pi}_c \)-Soft with \( \epsilon = 0.6 \) \( (\epsilon = 0.8) \).

### D.3 Full results with \( |\mathcal{D}| = 3,000 \)

Figure 5 shows the results in the helicopter environment with a small dataset \( |\mathcal{D}| = 3,000 \). We observe that the estimated policies have a performance even lower than in the experiment with \( |\mathcal{D}| = 10,000 \). Yet, the algorithm Soft-SPIBB still manages to improve upon the true baseline policy \( \pi_b \) with all the baselines policies, obtaining a mean performance significantly above the average performance of \( \pi_b \), and a 10%-quantile slightly lower than that of the true baseline when using the estimated policies.

Finally, we notice that RaMDP’s performance indicators dramatically plummet, even largely lower than the behavioural cloning policies.

![Figure 5: \( |\mathcal{D}| = 3,000 \). The green dashed line shows the average and the caps show the 10% and 90% percentile.](image)

### D.4 Hyper-parameter search

The hyper-parameter search reported in Figure 6 gives us extra insights on the behavior of the algorithms SPIBB and Soft-SPIBB using estimated baselines. We can notice that these algorithms do not have a high sensitivity to their hyper-parameters, since the performance is stable in a wide range of values, specially the Soft-SPIBB variations. We sometimes notice a tradeoff that has to be made between variance reduction and expectation maximization. We may also notice that lower values of \( N_\lambda \) may work better for the SPIBB algorithm with baseline estimates. The camera-ready will complete and update the concerned plots accordingly.
Figure 6: $|\mathcal{D}| = 10,000$ in the two first rows and $|\mathcal{D}| = 3,000$ in the two last rows. The green dashed line shows the average and the caps show the 10% and 90% percentile.