Low Data in Reinforcement Learning

Contributors:

- Damien Ernst (dernst@uliege.be),
- Arthur Louette (arthur.uliege@uliege.be),
- Raphaël Fonteneau (Raphael.Fonteneau@uliege.be)

February 2025

Low data

In this lesson, we are interested in settings where data is considered as rare or costly to generate. This is what we mean by ”low data”.

Examples: data obtained from clinical trials, data from a system for which no simulator nor accurate model is available, data associated with complex humain interaction.

Counterexamples: Atari (Mnih et al. [2015]), Go (Silver et al. [2017]), StarCraft II (Vinyals et al. [2019]) ...

In the following, we will more specifically study two cases: the $K-$armed bandit problem, and the batch-mode Reinforcement Learning.

Outline

The $K-$ armed bandits

Imagine that you enter a casino room. There are several (let us say $K \in \mathbb{N}$) slot machines that you can play. Playing any machine will provide you a random reward drawn according to an unknown probability distribution. Assuming that all the K probability distributions do not have the same expected value, then it means that, on average, at least one machine is more interesting to be played than the others. The question is: how to discover such a machine?

The $K-$armed bandit problem is a typical example of the so-called exploration vs exploitation dilemma.

Interesting fact: such a problem was not originally formalized to play casino games, but rather in the context of clinical trials (see Thompson [1933], Robbins [1952]). How to investigate the effects of different experimental treatments while minimizing patient losses?

The K-armed bandit

In the following, we mainly rely on the formalism used in Munos et al. [2014].

Consider K arms (actions, choices) defined by distributions $(\nu_k)_{1 \leq k \leq K}$ with bounded support $[0, 1]$ that are initially unknown to the decision maker (or the player).

At each round $t = 1, \dots, n$, the decision maker takes a decision by selecting an arm $a_t \in \{1, \dots, K\}$ and receives a reward $r_t \sim \nu_{a_t}$ which is a random sample drawn from the distribution ν_{a_t}, corresponding to the selected arm a_t and assumed to be independent of previously received rewards. The goal of the decision maker is to maximize the sum of obtained rewards in expectation.

In the following, we also denote by ρ_k the expected values of each arm, by ρ^* such best value and by k^* one best arm (there may exit several):

$$\begin{aligned} \rho_k &= \mathbb{E}[\nu_k] \\ \rho^* &= \arg\max_k \rho_k \\ &= \rho_{k^*} \end{aligned}$$

The K-armed bandit

If the arm distributions where known, the agent would select the arm with the highest mean at each time step and obtain an expected cumulative reward $n * \rho^*$ after n steps.

However, since the distributions of the arms are initially unknown, we need to pull each arm several times to progressively get information - a process called exploration - and while this knowledge about the arms improves, we may increasingly often pull the best arms - a process called exploitation. Balancing between these two processes is called the exploration-exploitation trade-off.

Note that balancing between exploration and exploitation may particularly make sense in contexts where data are costly and/or take time to generate.

The cumulative regret

In order to assess the performance of any pulling strategy, we compare its performance to an oracle strategy that would know the distribution in advance - and would thus be able to always play an optimal arm. For that purpose, we define the notion of cumulative regret:

$$\forall n \in \mathbb{N}, R_n \equiv n\rho^* - \sum_{t=1}^{n} r_t \tag{1}$$

The cumulative regret defines a loss, in terms of cumulative rewards, resulting from not having acess to the knowledge of the/a best distribution from the beginning. Therefore, it is interesting to design strategies with low cumulative regret.

In the following, we introduce the following notations: Δ_k is the positive gap between the expected value of arm k and the best expected value among all arms, and $T_k(n)$ is the number of pulls of arm k during the n first time steps:

$$\Delta_k \equiv \rho^* - \rho_k \tag{2}$$

$$T_k(n) \equiv \sum_{t=1}^{n} \mathbb{I}_{\{a_t = k\}} \tag{3}$$

Using the law of total expectation, we have:

$$\begin{aligned} \mathbb{E}\left[R_n\right] &= n\rho^* - \mathbb{E}\left[\sum_{t=1}^{n} \rho_{a_t}\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{K} T_k(n)\left(\rho^* - \rho_k\right)\right] \\ &= \sum_{k=1}^{K} \mathbb{E}\left[T_k(n)\right] \Delta_k \end{aligned} \tag{4}$$

Equation 4 expresses the idea that, while it is necessary to sample them to acquire information about the arms, sub-optipmal arms should not be played too often.

This problem has challenged many researchers, and several solutions have been proposed: bayesian exploration, ε-greedy exploration, soft-max exploration, follow the perturbed leader, optimistic exploration,.... In the following, we study more carefully an algorithm from the optimistic exploration family: the Upper Confidence Bounds (UCB) algorithm Auer et al. [2002].

The Upper Confidence Bound (UCB) algorithm

The Upper Confidence Bounds (UCB) stragegy consists of selecting at each time step t an arm with largest B-value:

$$\forall t \in \mathbb{N}^*, a_t \in \underset{k \in \{1,\ldots,K\}}{\arg\max}\ B_{t,T_k(t-1)}(k) \tag{5}$$

where, for each arm k, the B-value of an arm k is defined as:

$$\forall t \in \mathbb{N}^*, B_{t,s}(k) \equiv \hat{\rho}_{k,s} + \sqrt{C_{\text{UCB}} \frac{\log(t)}{s}} \tag{6}$$

$$\hat{\rho}_{k,s} \equiv \frac{1}{s} \sum_{i=1}^{s} X_{k,i} \tag{7}$$

where $X_{k,i}$ denotes the reward received when pulling arms k for the $i-$th time. Note that, if we denote by $\tau_{k,i}$ the random time corresponding to the instant when we pull the arm k for the $i - th$ time, we have:

$$X_{k,i} = r_{\tau_{k,i}} \tag{8}$$

In the original paper of Auer et al., the constant C_{UCB} is equal to $C_{\text{UCB}} = 2$, but in this course, we consider $C_{\text{UCB}} = \frac{3}{2}$.

Chernoff-Hoeffding inequalities

The UCB strategy can be considered as optimistic because it selects the optimal arm in the most favorable environments that are (in high probability) compatible with the observations.

Let us first recall the Chernoff-Hoeffding inequality.

Lemma

Let $Y_i \in [0,1]$ *be* s *independent copies of a random variable of mean* ρ *and* $\varepsilon > 0$. *Then,*

$$\mathbb{P}\left(\frac{1}{s}\sum_{i=1}^{s} Y_i - \rho \geq \varepsilon\right) \leq e^{-2s\varepsilon^2}$$

$$\mathbb{P}\left(\frac{1}{s}\sum_{i=1}^{s} Y_i - \rho \leq -\varepsilon\right) \leq e^{-2s\varepsilon^2}$$

Chernoff-Hoeffding inequalities applied to the B-values

Applying the Chernoff-Hoeffding inequalities to the rewards obtained from the arms, with $\varepsilon = \sqrt{\frac{3\log(t)}{2s}}$, we have:

$$\forall 1 \leq s \leq t, \mathbb{P}\left(\hat{\rho}_{k,s} + \sqrt{\frac{3\log(t)}{2s}} \leq \rho_k\right) \leq e^{-3\log(t)} = t^{-3} \tag{9}$$

and

$$\forall 1 \leq s \leq t, \mathbb{P}\left(\hat{\rho}_{k,s} - \sqrt{\frac{3\log(t)}{2s}} \geq \rho_k\right) \leq e^{-3\log(t)} = t^{-3} \tag{10}$$

Thus, the B-values $B_{t,s}(k)$ are high-probability upper confidence bounds on the mean-value ρ_k:

$$\forall s \in \{1, \ldots, t\}, \mathbb{P}\left(B_{t,s}(k) \geq \rho_k\right) \leq 1 - t^{-3} \tag{11}$$

In the following, we will deduce a bound on the expected number of plays of sub-optimal arms by noticing that, with high probability, the sub-optimal arms are not played whenever their UCB is below ρ^*.

Bound on the cumuative regret of UCB

Lemma

Each sub-optimal arm k is played in expectation at most:

$$\mathbb{E}\left[T_k(n)\right] \leq 6\frac{\log(n)}{\Delta_k^2} + \frac{\pi^2}{3} + 1 \qquad (12)$$

time. Thus, the cumulative regret of UCB is bounded as follows:

$$\begin{aligned} \mathbb{E}\left[R_n\right] &= \sum_{k=1}^{K} \Delta_k \mathbb{E}\left[T_k(n)\right] && (13) \\ &\leq 6 \sum_{k:\Delta_k>0} \frac{\log(n)}{\Delta_k} + K\left(\frac{\pi^2}{3} + 1\right) && (14) \end{aligned}$$

Proof. Assume that a sub-optimal arm k is pulled at time t. This means that its B-value is larger than the B-values of the other arms, in particular that of the optimal arm k^*:

$$\hat{\rho}_{k,T_k(t-1)} + \sqrt{\frac{3\log(t)}{2T_k(t-1)}} \quad \geq \quad \hat{\rho}_{k^*,T_{k^*}(t-1)} + \sqrt{\frac{3\log(t)}{2T_{k^*}(t-1)}} \tag{15}$$

Then, observe that if the previous equation stands, it means that:

- (i) either the empirical mean of the optimal arm is under-estimated and is not within its confidence interval:

$$\hat{\rho}_{k^*,T_{k^*}(t-1)} + \sqrt{\frac{3\log(t)}{2T_{k^*}(t-1)}} \leq \rho^* \tag{16}$$

- (ii) either the empirical mean of the arm k is over-estimated and is not within its confidence interval:

$$\hat{\rho}_{k,T_k(t-1)} > \rho_k + \sqrt{\frac{3\log(t)}{2T_k(t-1)}} \tag{17}$$

- or (iii) the value of Δ_k and $T_k(t-1)$ satisfy the following relationship:

$$\rho_k + 2\sqrt{\frac{3\log(t)}{2T_k(t-1)}} > \rho^* \implies T_k(t-1) < \frac{6\log(t-1)}{\Delta_k^2} \tag{18}$$

Indeed, assume that all equations 16, 17 and 18 are wrong. Then,

$$
\begin{aligned}
B_{k^*,T_{k*}(t-1)}(k^*) &> \rho^* && \text{(since we assume not(eq. 16))} \\
&= \rho_k + \Delta_k && \text{(by definition of } \Delta_k\text{)} \\
&\geq \rho_k + 2\sqrt{\frac{3\log(t)}{2T_k(t-1)}} && \text{(since we assume not(eq. 18))} \\
&\geq \hat{\rho}_{k,T_k(t-1)} + \sqrt{\frac{3\log(t)}{2T_k(t-1)}} && \text{(since we assume not (eq. 17))} \\
&= B_{k,T_k(t-1)}(k) && \text{(by definition of the B-value)}
\end{aligned}
$$

which contradicts the fact that arm k was selected. So, at least one of the three conditions expressed in (16), (17) and (18) is true. This also says that, whenever (18) does not hold, i.e., $T_k(t-1) \geq \frac{6\log(t)}{\Delta_k^2} + 1$, then either the arm k was not pulled at time t, either one of the two (16) or (17) holds.

Let us define $u \equiv \left\lfloor \frac{6\log(n)}{\Delta_k^2} \right\rfloor + 1$. Then:

$$T_k(n) = \sum_{t=1}^{n} \mathbb{I}_{\{a_t = k\}} \quad (19)$$

$$= \sum_{t=1}^{n} \mathbb{I}_{\{(16) \text{ or } (17) \text{ or } (18) \text{ holds}\}} \quad (20)$$

$$T_k(n) \leq u + \sum_{t=u+1}^{n} \mathbb{I}_{\{a_t = k; T_k(t) > u\}} \quad (21)$$

$$\leq u + \sum_{t=u+1}^{n} \mathbb{I}_{\{(16) \text{ or } (17) \text{ holds}\}} \quad (22)$$

Using the Chernoff-Hoeffding inequality, it is possible to bound, for any t, the probability that event (16) holds:

$$\mathbb{P}\left(\exists 1 \leq s \leq t, \hat{\rho}_{k^*,s} + \sqrt{\frac{3\log(t)}{2s}} < \rho^*\right) \leq \sum_{s=1}^{t} \frac{1}{t^3} \quad (23)$$

$$= \frac{1}{t^2} \quad (24)$$

The same reasoning can be applied to bound the probability that event (17) holds:

$$\mathbb{P}\left(\exists 1 \leq s \leq t, \rho_k + \sqrt{\frac{3\log(t)}{2s}} < \hat{\rho}_{k,s}\right) \leq \frac{1}{t^2} \tag{25}$$

Taking the expected value of equation (22)

$$\begin{aligned}
\mathbb{E}\left[T_k(n)\right] &\leq \mathbb{E}[u] + \sum_{t=u+1}^{n} \mathbb{E}\left[\mathbb{I}_{\{(16) \text{ or } (17) \text{ holds}\}}\right] && (26)\\
&= \frac{6\log(n)}{\Delta_k^2} + 1 + \sum_{t=u+1}^{n} \mathbb{P}(\ (16) \text{ or } (17) \text{ holds}) && (27)\\
&\leq \frac{6\log(n)}{\Delta_k^2} + 1 + \sum_{t=u+1}^{n} \mathbb{P}((16) \text{ holds}) + \mathbb{P}((17) \text{ holds}) && (28)\\
&\leq \frac{6\log(n)}{\Delta_k^2} + 1 + 2\sum_{t=u+1}^{n} \frac{1}{t^2} && (29)\\
&\leq \frac{6\log(n)}{\Delta_k^2} + \frac{\pi^2}{3} + 1 && (30)
\end{aligned}$$

which ends the proof of the first part of the lemma.

Back to the expression of the cumulative regret given in equation (4), we have:

$$\begin{aligned}
\mathbb{E}\left[R_n\right] &= \sum_{k=1}^{K} \Delta_k \mathbb{E}[T_k(n)] && (31)\\
&\leq \sum_{k:\Delta_k>0} \Delta_k \left(\frac{6\log(n)}{\Delta_k^2} + \frac{\pi^2}{3} + 1\right) && (32)\\
&= 6 \sum_{k:\Delta_k>0} \frac{\log(n)}{\Delta_k} + K\left(\frac{\pi^2}{3} + 1\right) && (33)
\end{aligned}$$

which ends the second part of the lemma.

Observe that this bound depends on one key characteristics of the distributions: the gaps Δ_k.

We now give a final result which bounds the the expected regret independently from the values of the gaps.

Corollary

The expected regret of UCB is bounded as:

$$\mathbb{E}\left[R_n\right] \leq \sqrt{Kn\left(6\log(n) + \frac{\pi^2}{3} + 1\right)} \tag{34}$$

Proof. Using the Cauchy-Schwarz inequality, we have:

$$\begin{aligned}
\mathbb{E}[R_n] &= \sum_{k=1}^{K} \sqrt{\Delta_k^2 \mathbb{E}[T_k(n)]}\sqrt{\mathbb{E}[T_k(n)]} && (35)\\
&\leq \sqrt{\left(\sum_{k=1}^{K} \Delta_k^2 \mathbb{E}[T_k(n)]\right)\left(\sum_{k=1}^{K} \mathbb{E}[T_k(n)]\right)} && (36)\\
&\leq \sqrt{\left(\sum_{k=1}^{K} \Delta_k^2 \left(\frac{6\log(n)}{\Delta_k^2} + \frac{\pi^2}{3} + 1\right)\right)\left(\sum_{k=1}^{K} \mathbb{E}[T_k(n)]\right)} && (37)
\end{aligned}$$

Observing that $\sum_{k=1}^{K} \mathbb{E}[T_k(n)] = \mathbb{E}[\sum_{k=1}^{K} T_k(n)] = \mathbb{E}[n] = n$, and remembering that $\Delta_k \leq 1$, we can further developp

$$\mathbb{E}[R_n] \leq \sqrt{\left(K \times 6\log(n) + \sum_{k=1}^{K} \Delta_k^2 \left(\frac{\pi^2}{3} + 1\right)\right)\left(n\right)} \tag{38}$$

$$\leq \sqrt{n\left(K \times 6\log(n) + \sum_{k=1}^{K} \left(\frac{\pi^2}{3} + 1\right)\right)} \tag{39}$$

$$\leq \sqrt{Kn\left(6\log(n) + \frac{\pi^2}{3} + 1\right)} \tag{40}$$

which ends the proof. ■

Illustration

We consider 3 arms, with $\rho_1 = 0.3, \rho_2 = 0.5, \rho_3 = 0.4$, with uniform distributions centered arnound the mean with width 0.3.

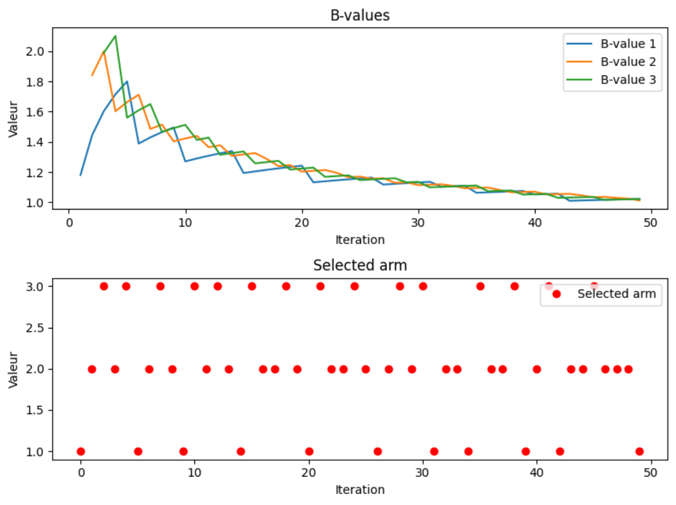


Figure 1: Playing the UCB algorithm

Synthesis

The K-armed bandit problem can be seen as a decision making problem where there is only 1 state: $\mathcal{S} = \{s\}$ and K discrete actions $\mathcal{A} = \{a_1, \ldots, a_K\}$ corresponding to each arm. The dynamics is trivial in the sense that $f(s, a, w) = s, \forall a \in \mathcal{A}, \forall w \sim P_w(\cdot)$.

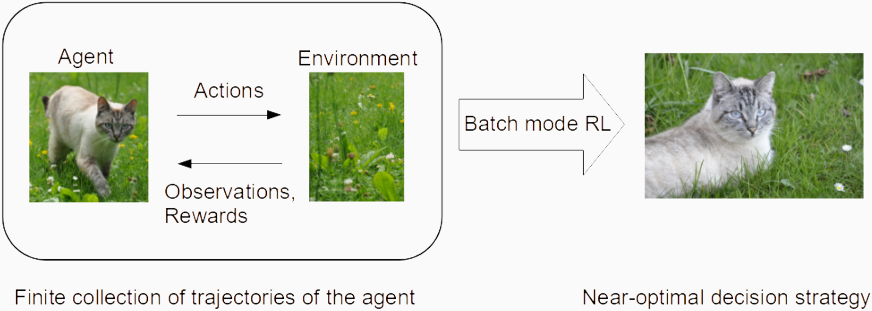
In this context, the UCB algorithm seeks to minimize the cumulative regret, i.e. tries to avoid taking non-optimal decisions too often.

The notion of regret has also been exploited in the context of classical MDPs. There is a variety of decision making algorithms for MDPs paying attention to their performance during learning.

Batch mode Reinforcement Learning

Batch mode Reinforcement Learning

Batch mode RL: all the available information is contained in a batch collection of data. Batch mode RL aims at computing a (near-)optimal policy from this collection of data



Low Data Batch-mode Reinfocement Learning

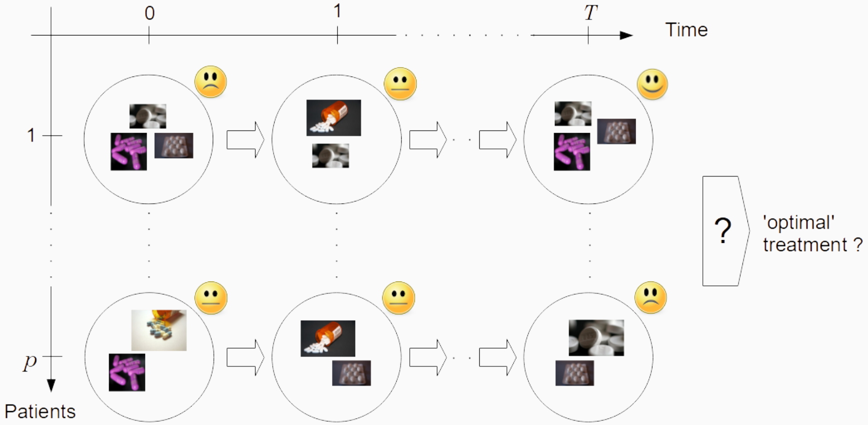
How low is low data?

Two examples:

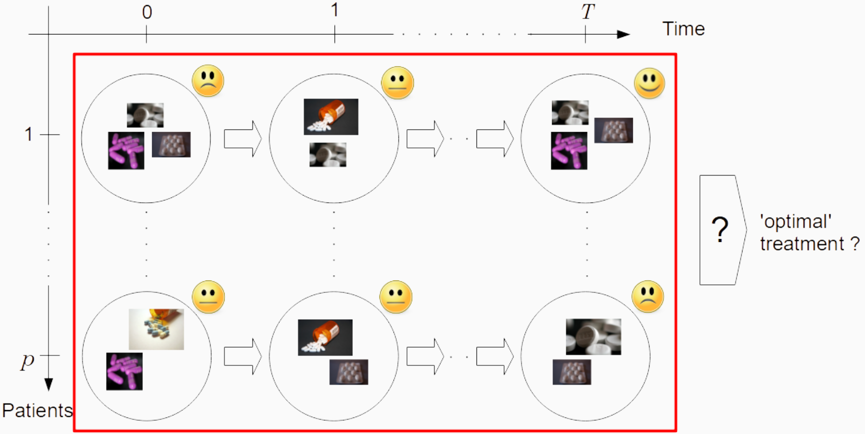
- Learning to play Atari Mnih et al. [2015] : tens of million of frames for solving Atari games - access to a model for performing simulations
- The STAR*D data set Nelson [2006] : tens of thousand multiple-item questionnaires for better treating depression (Dynamic Treatment Regimes) - no access to any reliable simulator

Example: Dynamic Treatment Regimes

"A dynamic treatment regime is a list of decision rules, one per time interval, for how the level of treatment will be tailored through time to an individual's changing status." Murphy [2003].



Example: Dynamic Treatment Regimes



Batch collection of trajectories of patients

DTR: objectives

Main goal: Finding a “good” policy

Many associated subgoals:

- Evaluating the performance of a given policy
- Computing performance guarantees
- Computing safe policies
- Choosing how to generate additional transitions ...

Main difficulties

Main difficulties of the batch mode setting:

- Dynamics and reward functions are unknown (and not accessible to simulation)
- The state-space and/or the action space are large or continuous
- The environment may be highly stochastic
- Data

Usual approach

To combine dynamic programming with function approximators (neural networks, regression trees, SVM, linear regression over basis functions, etc)

Function approximators have two main roles:

- To offer a concise representation of state-action value function for deriving value / policy iteration algorithms
- To generalize information contained in the finite sample

Remaining challenges

The black box nature of function approximators may have some unwanted effects:

- hazardous generalization
- difficulties to compute performance guarantees
- unefficient use of optimal trajectories

A proposition: synthesizing artificial trajectories.

Synthesizing Artificial Trajectories

Formalization

- Stochastic discrete-time systems:

$$s_{t+1} = f\left(s_t, a_t, w_t\right) \qquad \forall t \in \{0, \ldots, T-1\}$$

  where s_t belongs to a state space $\mathcal{S} \subset \mathbb{R}^d$, where $\mathbb{R}^d$ is the $d-$dimensional Euclidean space and $T \in \mathbb{N} \setminus \{0\}$ denotes the finite optimization horizon.
- At every time $t \in \{0, \ldots, T-1\}$, the system can be controlled by taking an action $a_t \in \mathcal{A}$, and is subject to a random disturbance $w_t \in \mathcal{W}$ drawn according to a probability distribution $P_w(\cdot)$. Here the fundamental assumption is that w_t is independent of $w_{t-1}, w_{t-2}, \ldots, w_0$ given s_t and a_t; to simplify all notations and derivations, we furthermore impose that the process is time-invariant and does not depend on the states and actions s_t, a_t.
- With each system transition from time t to $t+1$ is associated a reward signal:

$$r_t = r\left(s_t, a_t, w_t\right) \in \mathbb{R} \qquad \forall t \in \{0, \ldots, T-1\}\ .$$

Let $\pi : \{0, \ldots, T-1\} \times \mathcal{S} \to \mathcal{A}$ be a control policy. When starting from a given initial state s_0 and following the control policy π, an agent will get a random sum of rewards signal $R^{T\,\pi(s_0, w_0, \ldots, w_{T-1})}$:

$$\begin{aligned} R_T^\pi(s_0, w_0, \ldots, w_{T-1}) &= \sum_{t=0}^{T-1} r(s_t, \pi(t, s_t), w_t) \\ \text{with} \quad s_{t+1} &= f(s_t, \pi(t, s_t), w_t) \qquad \forall t \in \{0, \ldots, T-1\} \\ w_t &\sim P_w(\cdot)\ . \end{aligned}$$

Note : in the following, $R_T^\pi(s_0, w_0, \ldots, w_{T-1})$ may be simply denoted $R_T^\pi(s_0)$ in the proofs.

Formalization

In RL, the classical performance criterion for evaluating a policy π is its expected $T-$stage return:

Definition (Expected $T-$stage Return)

$$V_T^\pi(s_0) = \mathbb{E}\left[R_T^\pi(s_0, w_0, \ldots, w_{T-1})\right] ,$$

but, when searching for risk-aware policies, it is also of interest to consider a risk-sensitive criterion:

Definition (Risk-sensitive $T-$stage Return)

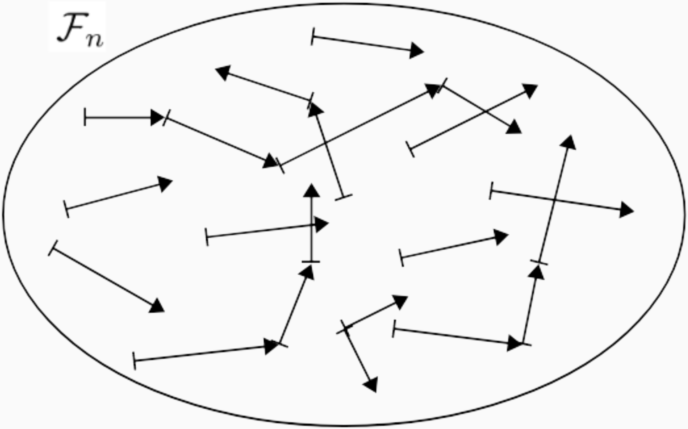
Let $b \in \mathbb{R}$ and $c \in [0,1[$.

$$V_{RS,T}^{\pi,(b,c)}(s_0) = \begin{cases} -\infty & \text{if } \mathbb{P}\left(R_T^\pi(s_0, w_0, \ldots, w_{T-1}) < b\right) > c , \\ V_T^\pi(s_0) & \text{otherwise} . \end{cases}$$

One-step system transitions

The system dynamics, reward function and disturbance probability distribution are unknown.

Instead, we have access to a sample of one-step system transitions:



Sample of one-step system transitions

Definition (Sample of transitions)

Let

$$\mathcal{P}_n = \left\{ \left(s^l, a^l \right) \right\}_{l=1}^{n} \in (\mathcal{S} \times \mathcal{A})^n$$

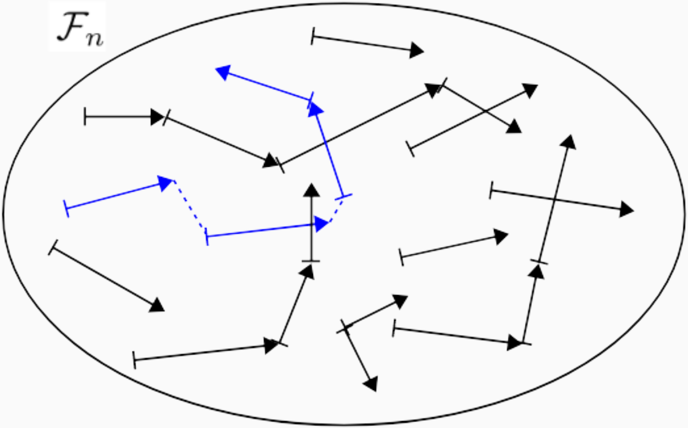
be a given set of state-action pairs. Consider the ensemble of samples of one-step transitions of size n that could be generated by complementing each pair (s^l, a^l) of $\mathcal{P}_n$ by drawing for each l a disturbance signal w^l at random from $P_w(\cdot)$, and by recording the resulting values of $r(s^l, a^l, w^l)$ and $f(s^l, a^l, w^l)$. We denote by $\tilde{\mathcal{F}}_n \left(\mathcal{P}_n, w^1, \ldots, w^n \right)$ one such "random" set of one-step transitions defined by a random draw of n i.i.d. disturbance signals $w^l, \quad l = 1 \ldots n$. We assume that we know one realization of the random set $\tilde{\mathcal{F}}_n \left(\mathcal{P}_n, w^1, \ldots, w^n \right)$, that we denote by $\mathcal{F}_n$:

$$\begin{aligned} \mathcal{F}_n &= \left\{ \left(s^l, a^l, r^l, s'^l \right) \right\}_{l=1}^{n} \\ \forall l \in \{1, \ldots, n\}, \qquad r^l &= r \left(s^l, a^l, w^l \right) , \\ y^l &= f \left(s^l, a^l, w^l \right) , \end{aligned}$$

for some realizations of the disturbance process $w^l \sim P_w(\cdot)$.

Artificial trajectories

Artificial trajectories are (ordered) sequences of elementary pieces of trajectories Fonteneau et al. [2013]:



Artificial trajectories: what for?

Artificial trajectories can help for:

- Estimating the performances of policies
- Computing performance guarantees
- Computing safe policies
- Choosing how to generate additional transitions

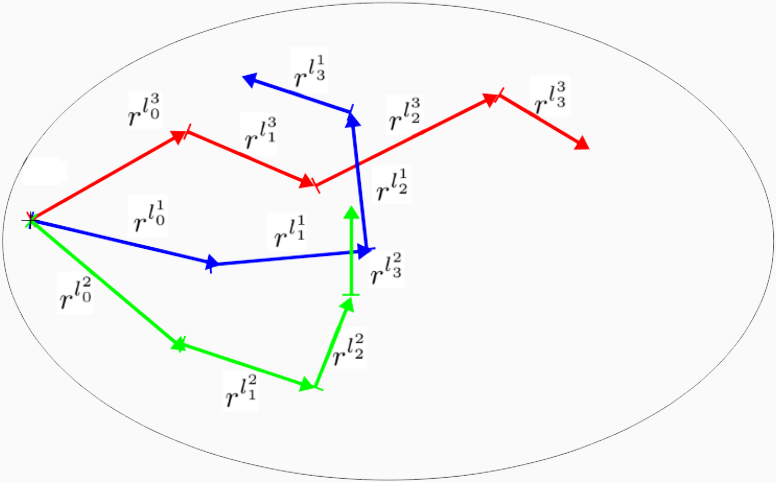
Artificial trajectories: what for?

Artificial trajectories can help for:

- **Estimating the performances of policies**
- Computing performance guarantees
- Computing safe policies
- Choosing how to generate additional transitions

Classical Monte Carlo Estimation

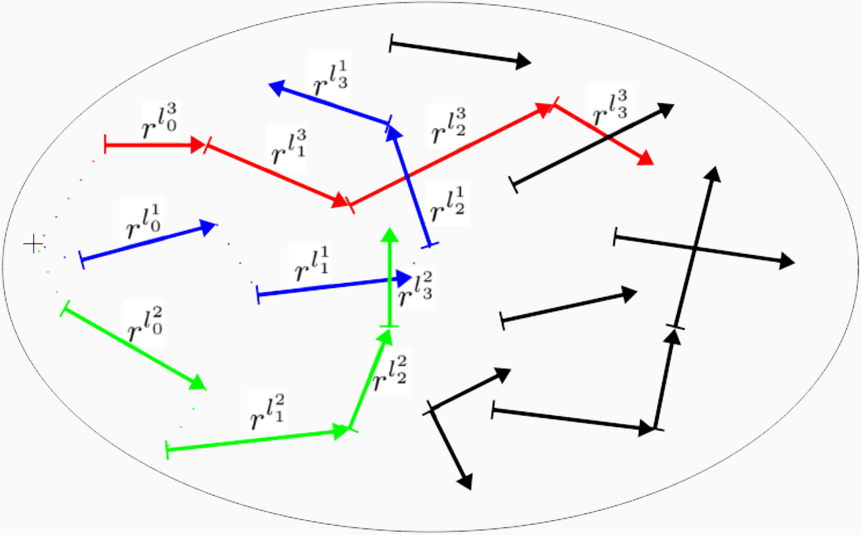
If the system dynamics and the reward function were accessible to simulation, then Monte Carlo estimation would allow estimating the performance of π:



Model-free Monte Carlo estimation

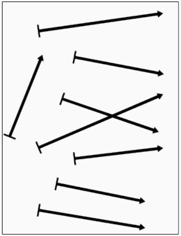
- We propose an approach that mimics MC estimation by rebuilding p artificial trajectories from one-step system transitions
- These artificial trajectories are built so as to minimize the discrepancy (using a distance metric Δ) with a classical MC sample that could be obtained by simulating the system with the policy π; each one step transition is used at most once
- We average the cumulated returns over the p artificial trajectories to obtain the Model-free Monte Carlo estimator (MFMC) of the expected return of π:

$$\mathfrak{M}^{\pi}_{T,p}\left(\mathcal{F}_n, s_0\right) = \frac{1}{p}\sum_{i=1}^{p}\sum_{t=0}^{T-1} r^{l_t^i}.$$


$r^{l_3^1}$
$r^{l_0^3}$
$r^{l_1^3}$
$r^{l_2^3}$
$r^{l_3^3}$
$r^{l_0^1}$
$r^{l_2^1}$
$r^{l_1^1}$
$r^{l_3^2}$
$r^{l_0^2}$
$r^{l_2^2}$
$r^{l_1^2}$

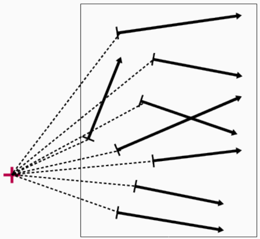
Model-free Monte Carlo estimation

Example with T = 3, p = 2, n = 8

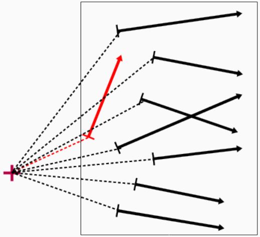


+

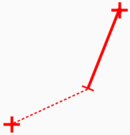
Model-free Monte Carlo estimation

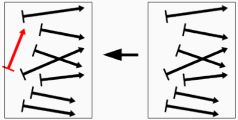


Model-free Monte Carlo estimation

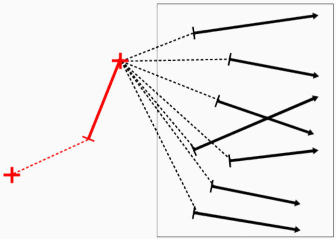


Model-free Monte Carlo estimation

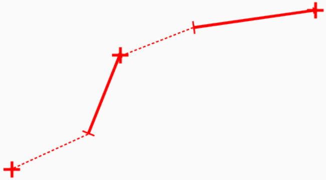


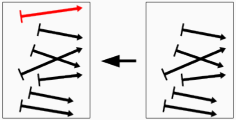


Model-free Monte Carlo estimation

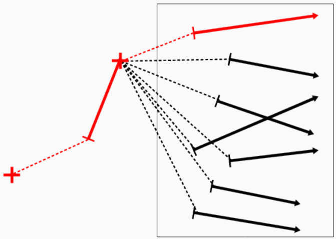


Model-free Monte Carlo estimation

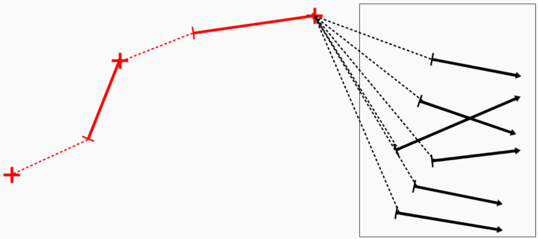




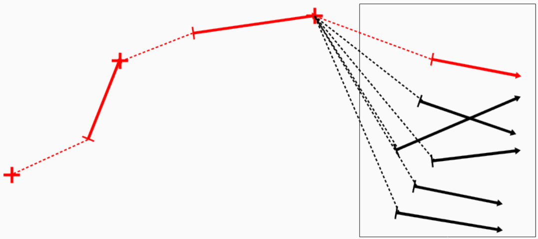
Model-free Monte Carlo estimation



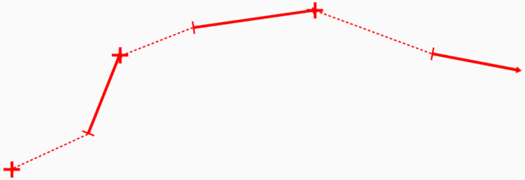
Model-free Monte Carlo estimation



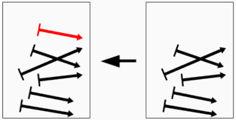
Model-free Monte Carlo estimation



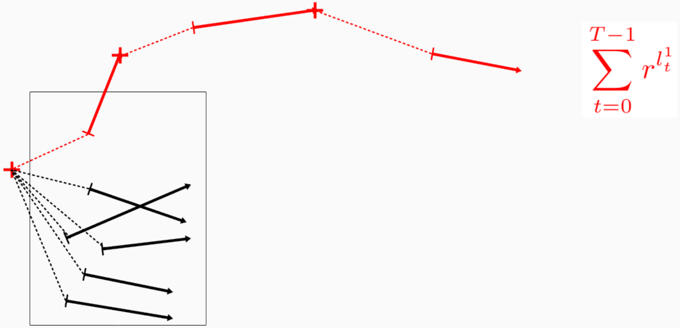
Model-free Monte Carlo estimation



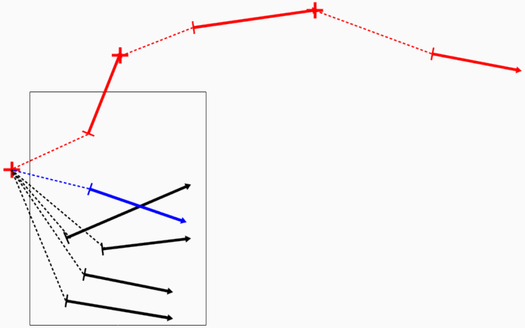
$$\sum_{t=0}^{T-1} r^{l_t^1}$$



Model-free Monte Carlo estimation

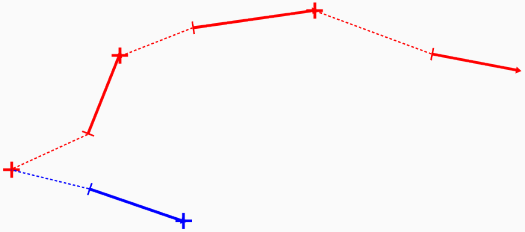


Model-free Monte Carlo estimation

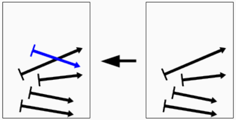


$$\sum_{t=0}^{T-1} r^{l_t^1}$$

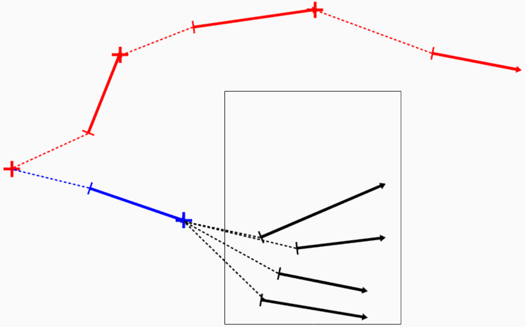
Model-free Monte Carlo estimation



$$\sum_{t=0}^{T-1} r^{l_t^1}$$

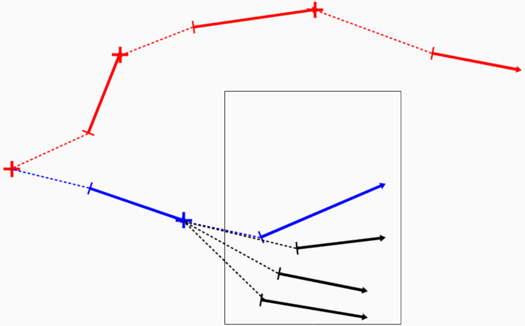


Model-free Monte Carlo estimation



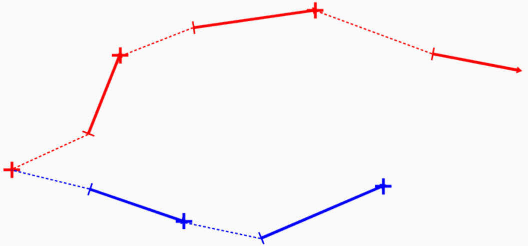
$$\sum_{t=0}^{T-1} r^{l_t^1}$$

Model-free Monte Carlo estimation

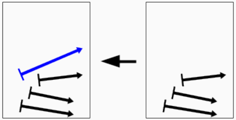


$$\sum_{t=0}^{T-1} r^{l_t^1}$$

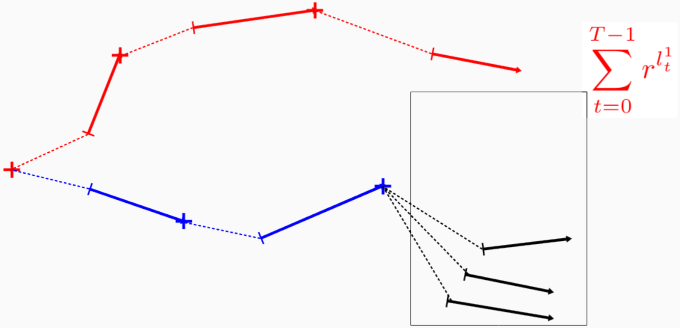
Model-free Monte Carlo estimation



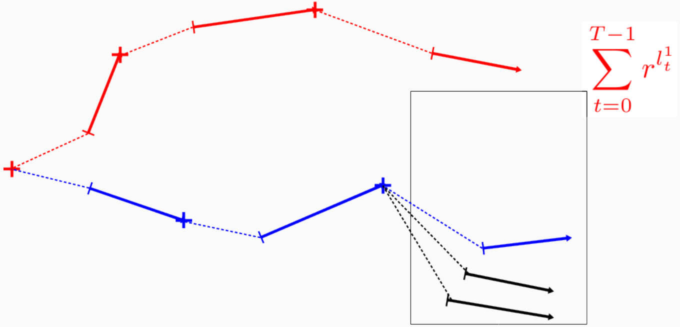
$$\sum_{t=0}^{T-1} r^{l_t^1}$$



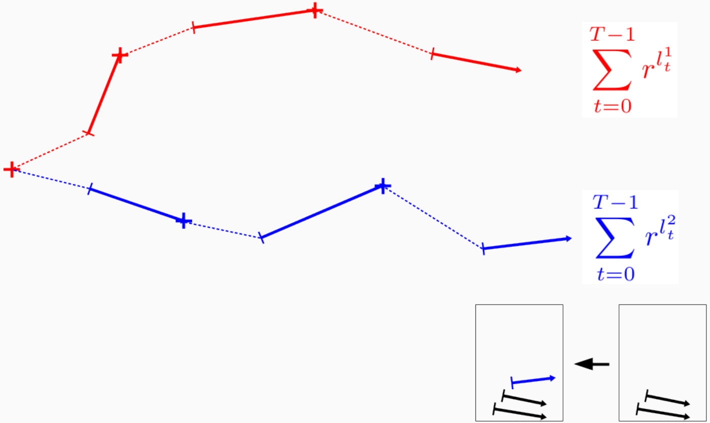
Model-free Monte Carlo estimation



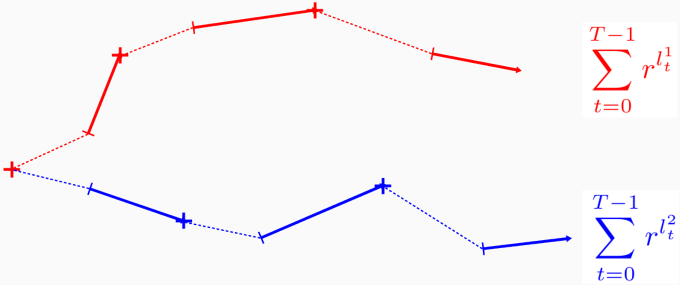
Model-free Monte Carlo estimation



Model-free Monte Carlo estimation



Model-free Monte Carlo estimation



Lipschitz continuity assumptions

Assumption: Lipschitz continuity of the functions f, r and π. We assume that the dynamics f, the reward function r and the policy π are Lipschitz continuous, i.e., we assume that there exist finite constants L_f, L_r and $L_\pi \in \mathbb{R}^+$ such that:
$\forall\ (s, s', a, a', w) \in \mathcal{S}^2 \times \mathcal{A}^2 \times \mathcal{W}$,

$$
\begin{aligned}
\|f(s,a,w) - f(s',a',w)\|_{\mathcal{S}} &\leq L_f(\|s - s'\|_{\mathcal{S}} + \|a - a'\|_{\mathcal{A}}),\\
|r(s,a,w) - r(s',a',w)| &\leq L_r(\|s - s'\|_{\mathcal{S}} + \|a - a'\|_{\mathcal{A}}),\\
\|\pi(t,s) - \pi(t,s')\|_{\mathcal{A}} &\leq L_\pi\|s - s'\|_{\mathcal{S}}\ , \forall t \in \{0, \ldots, T-1\}\ ,
\end{aligned}
$$

where $\|.\|_{\mathcal{S}}$ and $\|.\|_{\mathcal{A}}$ denote the chosen norms over the spaces $\mathcal{S}$ and $\mathcal{A}$, respectively.

Assumption: $\mathcal{S} \times \mathcal{A}$ is bounded. We suppose that $\mathcal{S} \times \mathcal{A}$ is bounded when measured using the distance metric Δ.

Definition (Distance Metric Δ)

$$\forall (s, s', a, a') \in \mathcal{S}^2 \times \mathcal{A}^2, \quad \Delta((s,a),(s',a')) = \|s - s'\|_{\mathcal{S}} + \|a - a'\|_{\mathcal{A}}.$$

Given $k \in \mathbb{N} \setminus \{0\}$ with $k \leq n$, we define the $k-$dispersion, $\alpha_k(\mathcal{P}_n)$ of the sample $\mathcal{P}_n$:

Definition ($k-$Dispersion)

$$\alpha_k(\mathcal{P}_n) = \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \Delta_k^{\mathcal{P}_n}(s,a) \ ,$$

where $\Delta_k^{\mathcal{P}_n}(s,a)$ denotes the distance of (s,a) to its $k-$th nearest neighbor (using the distance metric Δ) in the $\mathcal{P}_n$ sample. The $k-$dispersion is the smallest radius such that all Δ-balls in $\mathcal{S} \times \mathcal{A}$ of this radius contain at least k elements from $\mathcal{P}_n$; it can be interpreted as a worst-case measure on how closely $\mathcal{P}_n$ covers the $\mathcal{S} \times \mathcal{A}$ space using the k-th nearest neighbors.

$k-$dispersion

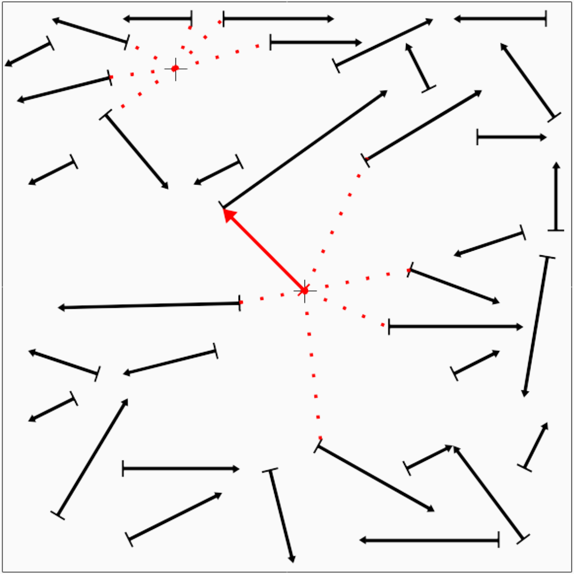


Figure 2: Schematic view of the $k-$dispersion

Expected value of the MFMC estimator

Definition (Expected Value of $\mathfrak{M}^{\pi}_{T,p}\left(\tilde{\mathcal{F}}_n\left(\mathcal{P}_n, w^1, \ldots, w^n\right), s_0\right)$)

We denote by $E^{\pi}_{T,p,\mathcal{P}_n}(s_0)$ the expected value:

$$E^{\pi}_{T,p,\mathcal{P}_n}(s_0) = \underset{w^1,\ldots,w^n \sim p_{\mathcal{W}}(.)}{\mathbb{E}}\left[\mathfrak{M}^{\pi}_{T,p}\left(\tilde{\mathcal{F}}_n\left(\mathcal{P}_n, w^1, \ldots, w^n\right), s_0\right)\right] .$$

Theorem (Bias Bound for $\mathfrak{M}^{\pi}_{T,p}\left(\tilde{\mathcal{F}}_n\left(\mathcal{P}_n, w^1, \ldots, w^n\right), s_0\right)$)

$$\left|V^{\pi}_{T}(s_0) - E^{\pi}_{T,p,\mathcal{P}_n}(s_0)\right| \leq C\alpha_{pT}\left(\mathcal{P}_n\right)$$

$$\text{with } C = L_r \sum_{t=0}^{T-1} \sum_{i=0}^{T-t-1} \left(L_f(1+L_\pi)\right)^i .$$

Proof. We first give three preliminary lemmas. Given a disturbance vector

$$\Omega = [\Omega(0), \dots, \Omega(T-1)] \in \mathcal{W}^T,$$

we define the Ω-disturbed state-action value function $Q^{\pi,\Omega}_{T-t}(s,a)$ for $t \in \{0, \dots, T-1\}$ as follows:

Definition (Ω-disturbed state-action value function)

$\forall t \in \{0, \dots, T-1\}, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \forall \Omega \in \mathcal{W}^T,$

$$Q^{\pi,\Omega}_{T-t}(s,a) = r(s,a,\Omega(t)) + \sum_{t'=t+1}^{T-1} r(s_{t'}, \pi(t', s_{t'}), \Omega(t'))$$

with $s_{t+1} = f(s,a,\Omega(t))$ and

$$\forall t' \in \{t+1, \dots, T-1\}\ , s_{t'+1} = f(s_{t'}, \pi(t', s_{t'}), \Omega(t')).$$

Expected value of the MFMC estimator

Then, we define the expected return given Ω the quantity

Definition (Expected return given Ω)

$\forall s_0 \in \mathcal{S}, \forall \Omega \in \mathcal{W}^T,$

$$\mathbb{E}[R_T^\pi(s_0)|\Omega] = \underset{w_0,\dots,w_{T-1}\sim p_{\mathcal{W}}(.)}{\mathbb{E}}[R_T^\pi(s_0)|w_0 = \Omega(0), \dots, w_{T-1} = \Omega(T-1)].$$

From there, we have the two following trivial results:

Lemma

$$\forall s_0 \in \mathcal{S}, \forall \Omega \in \mathcal{W}^T, \mathbb{E}[R_T^\pi(s_0)|\Omega] = Q_T^{\pi,\Omega}(s_0, h(0, s_0)) \ .$$

$\forall (s, a) \in \mathcal{S} \times \mathcal{A}, \ \forall \Omega \in \mathcal{W}^T,$

$$\begin{aligned} Q_{T-t+1}^{\pi,\Omega}(s,a) &= r(s,a,\Omega(t-1)) \\ &+ Q_{T-t}^{\pi,\Omega}\big(f\big(s,a,\Omega(t-1)\big), \pi\big(t, f(s,a,\Omega(t-1))\big)\big) \ . \end{aligned}$$

Then, we have the following lemma.

Lemma (Lipschitz Continuity of $Q^{\pi,\Omega}_{T-t}$)

$\forall t \in \{0, \ldots, T-1\}, \forall (s, s', a, a') \in \mathcal{S}^2 \times \mathcal{A}^2,$

$$\left| Q^{\pi,\Omega}_{T-t}(s,a) - Q^{\pi,\Omega}_{T-t}(s',a') \right| \leq L_{Q_{T-t}} \Delta((s,a),(s',a'))$$

with

$$L_{Q_{T-t}} = L_r \sum_{i=0}^{T-t-1} \left[L_f (1 + L_\pi) \right]^i .$$

Proof. We denote by $\mathcal{H}(T-t)$ the proposition:
$\mathcal{H}(T-t) : \forall (s, s', a, a') \in \mathcal{S}^2 \times \mathcal{A}^2,$

$$\left| Q^{\pi,\Omega}_{T-t}(s,a) - Q^{\pi,\Omega}_{T-t}(s',a') \right| \leq L_{Q_{T-t}} \Delta((s,a),(s',a')) \ .$$

We prove by induction that $\mathcal{H}(T-t)$ is true $\forall t \in \{0, \ldots, T-1\}$. For the sake of conciseness, we denote use the notation

$$\Delta^Q_{T-t} = \left| Q^{\pi,\Omega}_{T-t}(s,a) - Q^{\pi,\Omega}_{T-t}(s',a') \right| .$$

- **Basis:** $t = T-1$

We have

$$\Delta^Q_1 = |r(s,a,\Omega(T-1)) - r(s',a',\Omega(T-1)|,$$

and the Lipschitz continuity of r allows to write

$$\Delta^Q_1 \leq L_r \big(\|s - s'\|_{\mathcal{S}} + \|a - a'\|_{\mathcal{A}} \big) = L_r \Delta((s,a),(s',a')) .$$

This proves $\mathcal{H}(1)$.

Expected value of the MFMC estimator

- **Induction step:** We suppose that $\mathcal{H}(T-t)$ is true, $1 \leq t \leq T-1$.

One has

$$\begin{aligned}
\Delta^{Q}_{T-t+1} &= \left| Q^{\pi,\Omega}_{T-t+1}(s,a) - Q^{\pi,\Omega}_{T-t+1}(s',a') \right| \\
&= \Big| r(s,a,\Omega(t-1)) - r(s',a',\Omega(t-1)) \\
&+ Q^{\pi,\Omega}_{T-t}(f(s,a,\Omega(t-1)), \pi(t, f(s,a,\Omega(t-1)))) \\
&- Q^{\pi,\Omega}_{T-t}(f(s',a',\Omega(t-1)), \pi(t, f(s',a',\Omega(t-1)))) \Big|
\end{aligned}$$

and, from there,

$$\begin{aligned}
\Delta^{Q}_{T-t+1} &\leq \left| r(s,a,\Omega(t-1)) - r(s',a',\Omega(t-1)) \right| \\
&+ \Big| Q^{\pi,\Omega}_{T-t}(f(s,a,\Omega(t-1)), \pi(t, f(s,a,\Omega(t-1)))) \\
&- Q^{\pi,\Omega}_{T-t}(f(s',a',\Omega(t-1)), \pi(t, f(s',a',\Omega(t-1)))) \Big| .
\end{aligned}$$

Expected value of the MFMC estimator

$\mathcal{H}(T-t)$ and the Lipschitz continuity of r give

$$\begin{aligned}\Delta^Q_{T-t+1} &\leq L_r\Delta((s,a),(s',a'))\\ &+ L_{Q_{T-t}}\Delta((f(s,a,\Omega(t-1)),\pi(t,f(s,a,\Omega(t-1)))),\\ &\quad (f(s',a',\Omega(t-1)),\pi(t,f(s',a',\Omega(t-1))))) \; .\end{aligned}$$

Using the Lipschitz continuity of f and π, we have

$$\begin{aligned}\Delta^Q_{T-t+1} &\leq L_r\Delta((s,a),(s',a'))\\ &+ L_{Q_{T-t}}\big(L_f\Delta((s,a),(s',a')) + L_\pi L_f\Delta((s,a),(s',a'))\big),\end{aligned}$$

and, from there,

$$\Delta^Q_{T-t+1} \leq L_{Q_{T-t+1}}\Delta((s,a),(s',a'))$$

since

$$L_{Q_{T-t+1}} \doteq L_r + L_{Q_{T-t}}L_f(1+L_\pi).$$

This proves $\mathcal{H}\;(T-t+1)$ and ends the proof.

Expected value of the MFMC estimator

Definition (Disturbance vector associated with a broken trajectory)

Given a broken trajectory

$$\tau^i = \left[\left(s^{l_t^i}, a^{l_t^i}, r^{l_t^i}, s'^{l_t^i}\right)\right]_{t=0}^{T-1}$$

we denote by Ω^{τ^i} its associated disturbance vector

$$\Omega^{\tau^i} = \left[w^{l_0^i}, \ldots, w^{l_{T-1}^i}\right],$$

i.e. the vector made of the T unknown disturbances that affected the generation of the one-step transitions $\left(s^{l_t^i}, a^{l_t^i}, r^{l_t^i}, s'^{l_t^i}\right)$.

Expected value of the MFMC estimator

We give the following lemma.

Lemma (Bounds on the expected return given Ω)

$\forall s_0 \in \mathcal{S}, \forall i \in \{1, \ldots, p\}$,

$$b^\pi(\tau^i, s_0) \leq \mathbb{E}\left[R_T^\pi(s_0) | \Omega^{\tau^i}\right] \leq a^\pi(\tau^i, s_0) \ ,$$

with

$$b^\pi(\tau^i, s_0) = \sum_{t=0}^{T-1} \left[r^{l_t^i} - L_{Q_{T-t}} \delta_t^i\right] \ ,$$

$$a^\pi(\tau^i, s_0) = \sum_{t=0}^{T-1} \left[r^{l_t^i} + L_{Q_{T-t}} \delta_t^i\right] \ ,$$

$$\delta_t^i = \Delta\left(\left(s^{l_t^i}, a^{l_t^i}\right), \left(s'^{l_{t-1}^i}, \pi\left(t, s'^{l_{t-1}^i}\right)\right)\right) \ , \forall t \in \{0, \ldots, T-1\} \ ,$$

$$s'^{l_{-1}^i} = s_0, \forall i \in \{1, \ldots, p\}.$$

Proof. Let us first prove the lower bound. With $a_0 = \pi(0, s_0)$, the Lipschitz continuity of $Q_T^{\pi,\Omega^{\tau^i}}$ gives

$$\left| Q_T^{\pi,\Omega^{\tau^i}}(s_0, a_0) - Q_T^{\pi,\Omega^{\tau^i}}(s^{l_0^i}, a^{l_0^i}) \right| \leq L_{Q_T} \Delta\left((s_0, a_0), \left(s^{l_0^i}, a^{l_0^i}\right) \right) .$$

According to Proposition (13),

$$Q_T^{\pi,\Omega^{\tau^i}}(s_0, a_0) = \mathbb{E}\left[R_T^{\pi}(s_0) | \Omega^{\tau^i} \right] .$$

Thus,

$$\begin{aligned} \left| \mathbb{E}\left[R_T^{\pi}(s_0) | \Omega^{\tau^i} \right] - Q_T^{\pi,\Omega^{\tau^i}} \left(s^{l_0^i}, a^{l_0^i}\right) \right| &= \left| Q_T^{\pi,\Omega^{\tau^i}}(s_0, \pi(0, s_0)) - Q_T^{\pi,\Omega^{\tau^i}} \left(s^{l_0^i}, a^{l_0^i}\right) \right| \\ &\leq L_{Q_T} \Delta\left((s_0, \pi(0, s_0)), \left(s^{l_0^i}, a^{l_0^i}\right) \right) . \end{aligned}$$

Expected value of the MFMC estimator

It follows that

$$Q_T^{\pi,\Omega^{\tau^i}}\left(s^{l_0^i}, a^{l_0^i}\right) - L_{Q_T}\delta_0^i \leq \mathbb{E}\left[R_T^\pi(s_0)|\Omega^{\tau^i}\right] .$$

Using previous equations, we have

$$\begin{aligned} Q_T^{\pi,\Omega^{\tau^i}}\left(s^{l_0^i}, a^{l_0^i}\right) &= r\left(s^{l_0^i}, a^{l_0^i}, w^{l_0^i}\right) \\ &+ Q_{T-1}^{\pi,\Omega^{\tau^i}}\left(f\left(s^{l_0^i}, a^{l_0^i}, w^{l_0^i}\right), \pi\left(1, f\left(s^{l_0^i}, a^{l_0^i}, w^{l_0^i}\right)\right)\right) . \end{aligned}$$

By definition of Ω^{τ^i}, we have

$$r\left(s^{l_0^i}, a^{l_0^i}, w^{l_0^i}\right) = r^{l_0^i}$$

and

$$f\left(s^{l_0^i}, a^{l_0^i}, w^{l_0^i}\right) = s'^{l_0^i} .$$

From there

$$Q_T^{\pi,\Omega^{\tau^i}}\left(s^{l_0^i}, a^{l_0^i}\right) = r^{l_0^i} + Q_{T-1}^{\pi,\Omega^{\tau^i}}\left(s'^{l_0^i}, \pi\left(1, s'^{l_0^i}\right)\right) ,$$

and

$$Q_{T-1}^{\pi,\Omega^{\tau^i}}\left(s'^{l_0^i}, \pi\left(1, s'^{l_0^i}\right)\right) + r^{l_0^i} - L_{Q_T}\delta_0^i \leq \mathbb{E}\left[R_T^\pi(s_0)|\Omega^{\tau^i}\right] .$$

Expected value of the MFMC estimator

The Lipschitz continuity of $Q_{T-1}^{\pi,\Omega^{\tau^i}}$ gives

$$\begin{aligned}
&\left| Q_{T-1}^{\pi,\Omega^{\tau^i}}\left(y^{l_0^i}, \pi\left(1, s'^{l_0^i}\right)\right) - Q_{T-1}^{\pi,\Omega^{\tau^i}}\left(s^{l_1^i}, a^{l_1^i}\right)\right| \\
&\leq L_{Q_{T-1}} \Delta\left(\left(s'^{l_0^i}, \pi\left(1, s'^{l_0^i}\right)\right), \left(s^{l_1^i}, a^{l_1^i}\right)\right) \\
&= L_{Q_{T-1}} \delta_1^i,
\end{aligned}$$

which implies that

$$Q_{T-1}^{\pi,\Omega^{\tau^i}}\left(s^{l_1^i}, a^{l_1^i}\right) - L_{Q_{T-1}} \delta_1^i \leq Q_{T-1}^{\pi,\Omega^{\tau^i}}\left(s'^{l_0^i}, \pi\left(1, s'^{l_0^i}\right)\right) \;.$$

We therefore have

$$Q_{T-1}^{\pi,\Omega^{\tau^i}}\left(s^{l_1^i}, a^{l_1^i}\right) + r^{l_0^i} - L_{Q_T} \delta_0^i - L_{Q_{T-1}} \delta_1^i \leq \mathbb{E}\left[R_T^\pi(s_0) | \Omega^{\tau^i}\right].$$

The proof is completed by iterating this derivation. The upper bound is proved similarly.

Expected value of the MFMC estimator

We give a third lemma.

Lemma

$\forall s_0 \in \mathcal{S}, \forall i \in \{1, \ldots, p\}$,

$$a^{\pi}\left(\tau^{i}, s_0\right) - b^{\pi}\left(\tau^{i}, s_0\right) \leq 2C\alpha_{pT}\left(\mathcal{P}_n\right)$$

with

$$C = \sum_{t=0}^{T-1} L_{Q_{T-t}} .$$

Proof. By construction of the bounds, one has

$$a^{\pi}\left(\tau^{i}, s_0\right) - b^{\pi}\left(\tau^{i}, s_0\right) = \sum_{t=0}^{T-1} 2L_{Q_{T-t}}\delta_t^i \ .$$

The MFMC algorithm chooses $p \times T$ different one-step transitions to build the MFMC estimator by minimizing the distance $\Delta((s'^{l_{t-1}^i}, \pi(t, s'^{l_{t-1}^i})), (x^{l_t^i}, u^{l_t^i}))$, so by definition of the k-sparsity of the sample $\mathcal{P}_n$ with $k = pT$, one has

$$\begin{aligned} \delta_t^i &= \Delta\left(\left(s'^{l_{t-1}^i}, \pi\left(t, s'^{l_{t-1}^i}\right)\right), \left(x^{l_t^i}, u^{l_t^i}\right)\right) \\ &\leq \Delta_{pT}^{\mathcal{P}_n}\left(s'^{l_{t-1}^i}, \pi\left(t, s'^{l_{t-1}^i}\right)\right) \\ &\leq \alpha_{pT}\left(\mathcal{P}_n\right) \ , \end{aligned}$$

which ends the proof.

Expected value of the MFMC estimator

Using those three lemmas, one can now compute an upper bound on the bias of the MFMC estimator.

Proof of Theorem By definition of $a^{\pi}(\tau^i, s_0)$ and $b^{\pi}(\tau^i, s_0)$, we have

$$\forall i \in \{1, \ldots, p\}, \frac{b^{\pi}\left(\tau^i, s_0\right) + a^{h}\left(\tau^i, s_0\right)}{2} = \sum_{t=0}^{T-1} r^{l_t^i} \ .$$

Then, according to Lemmas 16 and 17, we have $\forall i \in \{1, \ldots, p\}$,

$$\begin{aligned}
&\left| \underset{w^1,\ldots,w^n \sim P_w(.)}{\mathbb{E}} \left[\mathbb{E}\left[R_T^h(s_0) | \Omega^{\tau^i} \right] - \sum_{t=0}^{T-1} r^{l_t^i} \right] \right| \\
&\leq \underset{w^1,\ldots,w^n \sim P_w(.)}{\mathbb{E}} \left[\left| \mathbb{E}\left[R_T^{\pi}(s_0) | \Omega^{\tau^i} \right] - \sum_{t=0}^{T-1} r^{l_t^i} \right| \right] \\
&\leq C \alpha_{pT}(\mathcal{P}_n) \ .
\end{aligned}$$

Expected value of the MFMC estimator

Thus,

$$\left| \frac{1}{p} \sum_{i=1}^{p} \mathop{\mathbb{E}}_{w^1,\ldots,w^n \sim P_w(.)} \left[\mathbb{E}\left[R_T^\pi(s_0) | \Omega^{\tau^i} \right] - \sum_{t=0}^{T-1} r^{l_t^i} \right] \right|$$
$$\leq \frac{1}{p} \sum_{i=1}^{p} \left| \mathop{\mathbb{E}}_{w^1,\ldots,w^n \sim P_w(.)} \left[\mathbb{E}\left[R_T^\pi(s_0) | \Omega^{\tau^i} \right] - \sum_{t=0}^{T-1} r^{l_t^i} \right] \right|$$
$$\leq C \alpha_{pT} \left(\mathcal{P}_n \right) \ ,$$

which can be reformulated

$$\left| \mathop{\mathbb{E}}_{w^1,\ldots,w^n \sim P_w(.)} \left[\frac{1}{p} \sum_{i=1}^{p} \mathbb{E}\left[R_T^\pi(s_0) | \Omega^{\tau^i} \right] \right] - E_{T,p,\mathcal{P}_n}^\pi(s_0) \right| \leq C \alpha_{pT} \left(\mathcal{P}_n \right) \ ,$$

since

$$\frac{1}{p} \sum_{i=1}^{p} \sum_{t=0}^{T-1} r^{l_t^i} = \mathfrak{M}_{T,p}^\pi(\tilde{\mathcal{F}}_n, s_0) \ .$$

Expected value of the MFMC estimator

Since the MFMC algorithm chooses $p \times T$ different one-step transitions, all the disturbances $\left\{w^{l_t^i}\right\}_{i=1,t=0}^{i=p,t=T-1}$ are i.i.d. according to $P_w(.)$. For all $i \in \{1, \ldots, p\}$, The law of total expectation gives

$$\begin{aligned}
&\underset{w^{l_0^i},\ldots,w^{l_{T-1}^i} \sim P_w(.)}{\mathbb{E}}\left[\underset{w^{l_0^i},\ldots,w^{l_{T-1}^i} \sim P_w(.)}{\mathbb{E}}\left[R_T^\pi(s_0)|\Omega^{\tau^i}\right]\right] \\
&= \underset{w_0,\ldots,w_{T-1} \sim P_w(.)}{\mathbb{E}}\left[R_T^\pi(s_0)\right] \\
&= V_T^\pi(s_0)\ .
\end{aligned}$$

This ends the proof.

This formula shows that the bias is bounded closer to the target estimate if the sample sparsity is small. Note that the sample sparsity itself actually only depends on the sample $\mathcal{P}_n$ and on the value of p (it will increase with the number of trajectories used by our algorithm).

Variance of the MFMC estimator

We denote by $Var^{\pi}_{T,p,\mathcal{P}_n}(s_0)$ the variance of the MFMC estimator defined as follows.

Definition (Variance of the MFMC estimator)

$\forall s_0 \in \mathcal{S}$,

$$\begin{aligned} Var^{\pi}_{T,p,\mathcal{P}_n}(s_0) &= \underset{w^1,\ldots,w^n \sim P_w(.)}{Var} \left[\mathfrak{M}^{\pi}_{T,p}(\tilde{\mathcal{F}}_n, s_0)\right] \\ &= \underset{w^1,\ldots,w^n \sim P_w(.)}{\mathbb{E}} \left[\left(\mathfrak{M}^{\pi}_{T,p}\left(\tilde{\mathcal{F}}_n, s_0\right) - E^{\pi}_{T,p,\mathcal{P}_n}(s_0)\right)^2\right] . \end{aligned}$$

We give the following theorem.

Theorem (Variance of the MFMC estimator)

$\forall s_0 \in \mathcal{S}$,

$$Var^{\pi}_{T,p,\mathcal{P}_n}(s_0) \leq \left(\frac{\sigma_{R^{\pi}_T}(s_0)}{\sqrt{p}} + 2C\alpha_{pT}(\mathcal{P}_n)\right)^2$$

with $C = L_r \sum_{t=0}^{T-1} \sum_{i=0}^{T-t-1} [L_f(1+L_\pi)]^i$.

Variance of the MFMC estimator

Proof : your turn!

Or see Fonteneau et al. [2010a] or Fonteneau [2011].

Experimental illustration

Consider the system dynamics and the reward function given by

$$s_{t+1} = \sin\left(\frac{\pi}{2}(s_t + a_t + w_t)\right)$$

and

$$r(s_t, a_t, w_t) = \frac{1}{2\pi} e^{-\frac{1}{2}(s_t^2 + a_t^2)} + w_t$$

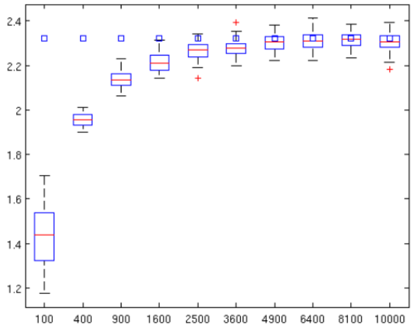
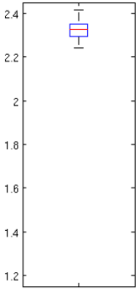
with the state space $\mathcal{X}$ being equal to $[-1, 1]$ and the action space $\mathcal{A}$ to $[-1, 1]$. The disturbance w_t is an element of the interval $\mathcal{W} = [-\frac{\varepsilon}{2}, \frac{\varepsilon}{2}]$ with $\varepsilon = 0.1$ and $p_{\mathcal{W}}$ is a uniform probability distribution over this interval. The optimization horizon T is equal to 15. The policy π whose performances have to be evaluated is

$$\pi(t, s) = -\frac{s}{2}, \qquad \forall s \in \mathcal{S}, \forall t \in \{0, \ldots, T-1\} \ .$$

The initial state of the system is set at $s_0 = -0.5$.

Influence of *n*

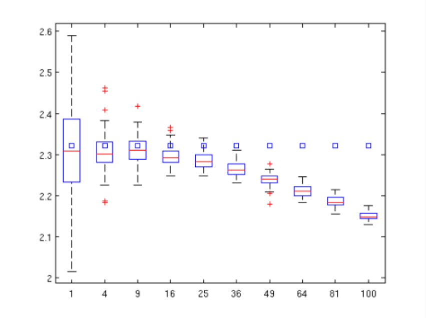
Simulations for p = 10, n = 100 … 10 000, uniform grid, T = 15, x_0 = - 0.5

Model-free Monte Carlo estimator	Monte Carlo estimator
 n = 100 … 10 000, p = 10	 p = 10

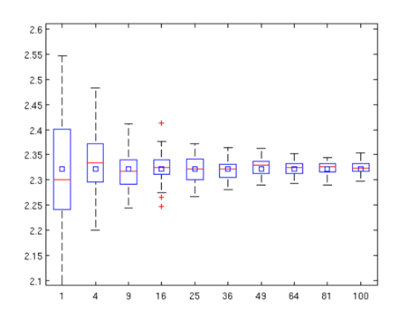
Influence of *p*

Simulations for p = 1 … 100, n = 10 000, uniform grid, T = 15, x_0 = - 0.5

Model-free Monte Carlo estimator	Monte Carlo estimator



p = 1 … 100, n=10 000



p = 1 … 100

Comparing MFMC with classical RL : let us define the finite horizon FQI iteration algorithm for policy evaluation (FQI-PE) that works by recursively computing a sequence of functions $\left(\hat{Q}^{\pi}_{T-t}(.\ ,.)\right)_{t=0}^{T-1}$ as follows:

Definition (FQI-PE Algorithm)

- $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.

$$\hat{Q}^{\pi}_{0}(s,a) = 0 \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \ ,$$

- For $t = T-1 \ldots 0$, build the dataset $D = \left\{\left(i^l, o^l\right)\right\}_{l=1}^{n}$:

$$\begin{aligned} i^l &= \left(s^l, a^l\right) \\ o^l &= r^l + \hat{Q}^{\pi}_{T-t-1}\left(s'^l, \pi(t+1, s'^l)\right) \end{aligned}$$

and use a regression algorithm $\mathcal{RA}$ to infer from D the function $\hat{Q}^{\pi}_{T-t}$:

$$\hat{Q}^{\pi}_{T-t} = \mathcal{RA}(D) \ .$$

The FQI -PE estimator of the policy π is given by:

Definition (FQI Estimator)

$$\hat{V}^{\pi}_{T,FQI}(\mathcal{F}_n, s_0) = \hat{Q}^{\pi}_{T}(s_0, \pi(0, s_0)) \ .$$

Experimental illustration

We propose to use a $k-$Nearest Neighbor algorithm ($k-$NN) as regression algorithm $\mathcal{RA}$. In the following, for a given state action couple $(s, a) \in \mathcal{S} \times \mathcal{A}$, we denote by $l_i(s, a)$ the lowest index in $\mathcal{F}_n$ of the i-th nearest one step transition from the state-action couple (s, a) using the distance measure Δ. The $k-$NN based FQI-PE estimation of π writes :

Definition ($k-$NN FQI-PE Algorithm)

- $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$,

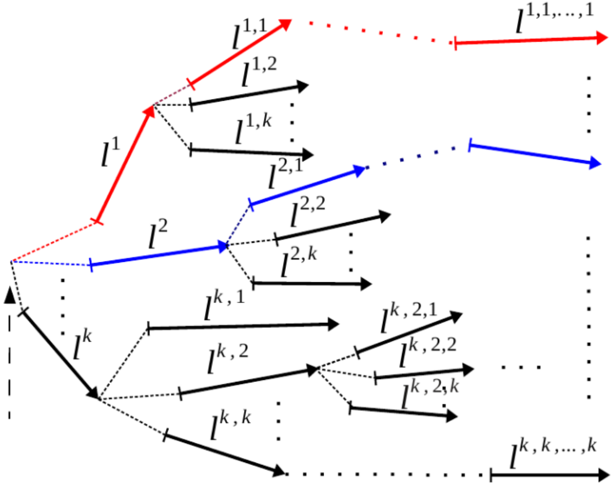
$$\hat{Q}_0^\pi(s, a) = 0 \ ,$$

- For $t = T - 1 \ldots 0 \ , \forall (s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\hat{Q}^\pi_{T-t}(s, a) = \frac{1}{k} \sum_{i=1}^{k} \left(r^{l_i(s,a)} + \hat{Q}^\pi_{T-t-1}\left(s'^{l_i(s,a)}, \pi\left(t+1, s'^{l_i(s,a)} \right) \right) \right) \ .$$

The $k-$NN FQI-PE estimator of the policy π is given by:

$$\hat{V}^\pi_{T,FQI}(\mathcal{F}_n, s_0) = \hat{Q}^\pi_T(s_0, \pi(0, s_0)) \ .$$


$l^{1,1}$
$l^{1,1,\dots,1}$
$l^{1,2}$
$l^{1,k}$
l^1
$l^{2,1}$
$l^{2,2}$
l^2
$l^{2,k}$
$l^{k,1}$
$l^{k,2,1}$
l^k
$l^{k,2}$
$l^{k,2,2}$
$l^{k,2,k}$
$l^{k,k}$
$l^{k,k,\dots,k}$

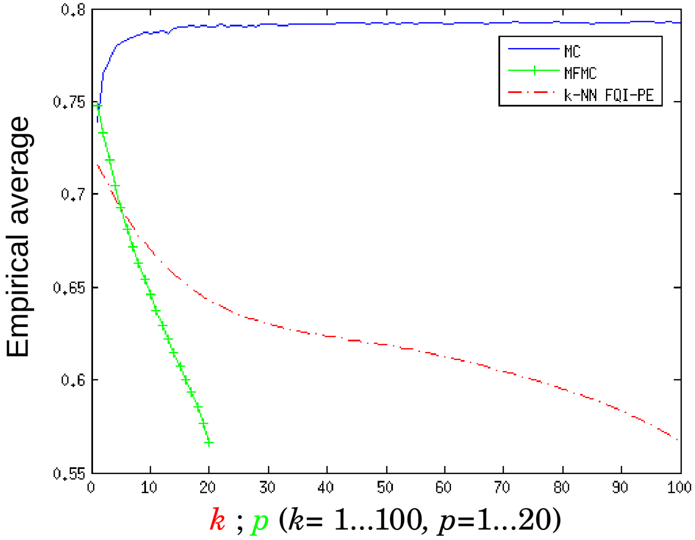


Figure 4: Comparison with the FQI-PE algorithm using k-NN, $n = 100$, $T = 5$.

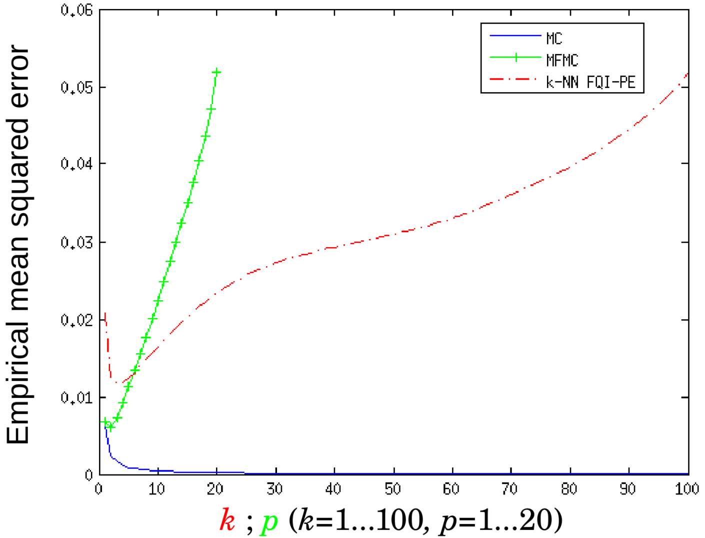
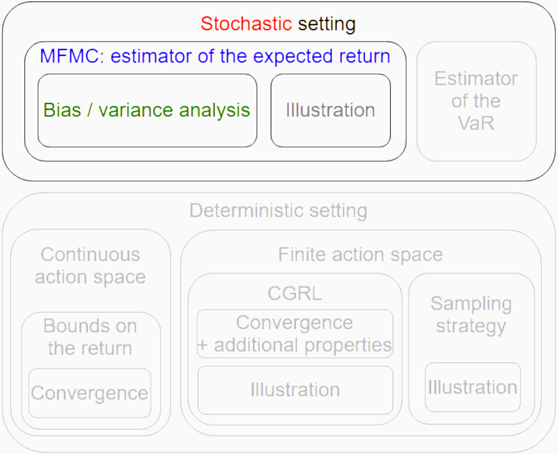


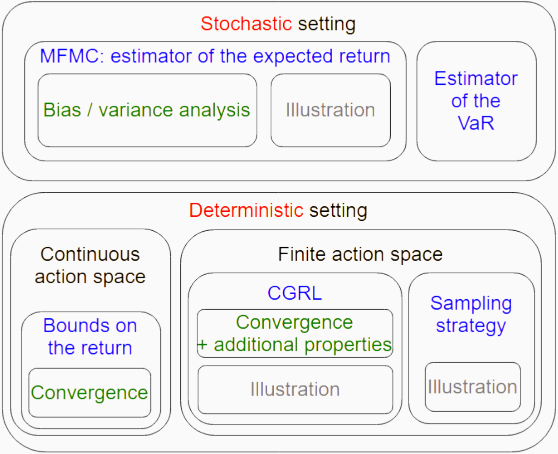
Figure 5: Comparison with the FQI-PE algorithm using k-NN, $n = 100$, $T = 5$.

Other bonuses

Other bonuses



Other bonuses



Risk-sensitive MFMC estimation

If we consider the p artificial trajectories that are rebuilt by the MFMC estimator, the risk-sensitive $T-$stage return $V^{\pi,(b,c)}_{T,RS}(s_0)$ can be efficiently approximated by the value $\tilde{V}^{\pi,(b,c)}_{T,RS}(s_0)$ defined as follows:

Definition (Estimate of the Risk-sensitive $T-$stage Return)

Let $b \in \mathbb{R}$ and $c \in [0,1[$.

$$\tilde{V}^{\pi,(b,c)}_{T,RS}(s_0) = \begin{cases} -\infty & \text{if } \frac{1}{p}\sum_{i=1}^{p} \mathbb{I}_{\{\mathbf{r}^i < b\}} > c\ , \\ \mathfrak{M}^{\pi}_{T}\left(\mathcal{F}_n, s_0\right) & \text{otherwise} \end{cases}$$

where $\mathbf{r}^i$ denotes the return of the $i-$th artificial trajectory:

$$\mathbf{r}^i = \sum_{t=0}^{T-1} r^{l^i_t}\ .$$

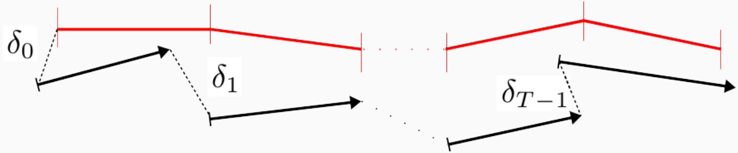
Lower and upper bounds in the deterministic case

From now, we assume a deterministic environment. More formally, we assume that the disturbances space is reduced to a single element $\mathcal{W} = \{0\}$ which concentrates on the whole probability mass $P_w(0) = 1$. We use the convention:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \qquad \begin{aligned} f(s,a) &= f(s,a,0) \ , \\ r(s,a) &= r(s,a,0) \ . \end{aligned}$$

We still assume that the functions f, r and π are Lipschitz continuous. Observe that, in a deterministic context, only one trajectory is needed to compute $V_T^{\pi}(s_0)$ by Monte Carlo estimation.

Lower and upper bounds in the deterministic case



Lower and upper bounds in the deterministic case

Lemma (Lower Bound from the MFMC)

Let $\left[\left(s^{l_t}, a^{l_t}, r^{l_t}, s'^{l_t}\right)\right]_{t=0}^{T-1}$ *be an artificial trajectory rebuilt by the MFMC algorithm when using the distance measure* Δ*. Then, we have*

$$\left|\mathfrak{M}^{\pi}_{T,1}(\mathcal{F}_n, s_0) - V^{\pi}_T(s_0)\right| \leq \sum_{t=0}^{T-1} L_{Q_{T-t}} \Delta\left((s'^{l_{t-1}}, h(t, s'^{l_{t-1}})), (s^{l_t}, a^{l_t})\right)$$

where

$$L_{Q_{T-t}} = L_r \sum_{i=0}^{T-t-1} \left(L_f\left(1 + L_\pi\right)\right)^i$$

and $s'^{l_{-1}} = s_0$.

The proof of this theorem can be found in Fonteneau et al. [2009].

Lower and upper bounds in the deterministic case

Since the previous result is valid for any artificial trajectory, we have:

Corollary (Lower Bound from any Artificial Trajectory)

Let $\left[\left(s^{l_t}, a^{l_t}, r^{l_t}, s'^{l_t}\right)\right]_{t=0}^{T-1}$ *be any artificial trajectory. Then,*

$$V_T^{\pi}(s_0) \geq \sum_{t=0}^{T-1} r^{l_t} - \sum_{t=0}^{T-1} L_{Q_{T-t}} \Delta\left(\left(s'^{l_{t-1}}, \pi(t, s'^{l_{t-1}})\right), \left(s^{l_t}, a^{l_t}\right)\right)$$

This suggests to identify an artificial trajectory that leads to the maximization of the previous lower bound:

Definition (Maximal Lower Bound)

$$\begin{aligned} L^{\pi}(\mathcal{F}_n, s_0) \quad = \quad & \max_{\left[\left(s^{l_t}, a^{l_t}, r^{l_t}, s'^{l_t}\right)\right]_{t=0}^{T-1} \in \mathcal{F}_n^T} \sum_{t=0}^{T-1} r^{l_t} \\ & - \sum_{t=0}^{T-1} L_{Q_{T-t}} \Delta\left(\left(s'^{l_{t-1}}, \pi(t, s'^{l_{t-1}})\right), \left(s^{l_t}, a^{l_t}\right)\right) . \end{aligned}$$

Lower and upper bounds in the deterministic case

Note that in the same way, a minimal upper bound can be computed:

Definition (Minimal Upper Bound)

$$U^{\pi}(\mathcal{F}_n, s_0) = \min_{[(s^{l_t}, a^{l_t}, r^{l_t}, s'^{l_t})]_{t=0}^{T-1} \in \mathcal{F}_n^T} \sum_{t=0}^{T-1} r^{l_t} + \sum_{t=0}^{T-1} L_{Q_{T-t}} \Delta\left((s'^{l_{t-1}}, \pi(t, s'^{l_{t-1}})), (s^{l_t}, a^{l_t})\right) .$$

Additionaly, we can prove that both the lower and the upper bound are tight, in the sense that they both converge towards $V_T^{\pi}(s_0)$ when the dispersion of the sample of system transitions $\mathcal{F}_n$ decreases towards zero.

Lemma (Tightness of the Bounds)

$$\exists C_b > 0 : \quad V_T^{\pi}(s_0) - L^{\pi}(\mathcal{F}_n, s_0) \leq C_b \alpha_1(\mathcal{P}_n)$$
$$U^{\pi}(\mathcal{F}_n, s_0) - V_T^{\pi}(s_0) \leq C_b \alpha_1(\mathcal{P}_n)$$

where $\alpha_1(\mathcal{P}_n)$ denotes the $1-$dispersion of the sample of system transitions $\mathcal{F}_n$.

Inferring safe policies from lower bounds

The previous results can be extended to the case where the action space $\mathcal{A}$ is finite (and thus discrete) by considering policies that are fully defined by a sequence of actions. Such policies can be qualified as "open-loop". Let $\mathcal{O}$ be the set of open-loop policies:

Definition (Open-loop Policies)

$$\mathcal{O} = \{o : \{0, \dots, T-1\} \to \mathcal{A}\}$$

Given an open-loop policy o, the (deterministic) $T-$stage return of o writes:

$$V_T^o(s_0) = \sum_{t=0}^{T-1} r(s_t, o(t))$$

with

$$s_{t+1} = f(s_t, o(t)), \quad \forall t \in \{0, \dots, T-1\}.$$

In the context of a finite action space, the Lipschitz continuity of f and r is:
$\forall\ (s, s', a) \in \mathcal{S}^2 \times \mathcal{A}$,

$$\begin{aligned} \|f(s,a) - f(s',a)\|_{\mathcal{S}} &\leq L_f \|s - s'\|_{\mathcal{S}}, \\ |r(s,a) - r(s',a)| &\leq L_r \|s - s'\|_{\mathcal{S}} \ . \end{aligned}$$

Since the action space is not normed anymore, we also need to redefine the sample dispersion.

Definition (Sample Dispersion)

We assume that the state space is bounded, and we define the sample dispersion $\alpha^*(\mathcal{P}_n)$ as follows:

$$\alpha^*(\mathcal{P}_n) = \sup_{s \in \mathcal{S}} \min_{l \in \{1,\ldots,n\}} \left\| s^l - s \right\|_{\mathcal{S}} \ .$$

Inferring safe policies from lower bounds

Let $o \in \mathcal{O}$ be an open-loop policy. We have the following result:

Lemma (Lower Bound - Open-loop Policy o)

Let $\left[\left(s^{l_t}, a^{l_t}, r^{l_t}, s'^{l_t}\right)\right]_{t=0}^{T-1}$ *be an artificial trajectory such that*

$$a^{l_t} = o(t) \quad \forall t \in \{0, \ldots, T-1\} \; .$$

Then,

$$V_T^o(s_0) \geq \sum_{t=0}^{T-1} r^{l_t} - \sum_{t=0}^{T-1} L'_{Q_{T-t}} \left\| s'^{l_{t-1}} - s^{l_t} \right\|_{\mathcal{S}} \; .$$

where

$$L'_{Q_{T-t}} = L_r \sum_{i=0}^{T-t-1} \left(L_f\right)^i \; .$$

Inferring safe policies from lower bounds

A maximal lower bound can then be computed by maximizing the previous bound over the set of all possible artificial trajectories that satisfy the condition $a^{l_t} = o(t) \quad \forall t \in \{0, \ldots, T-1\}$. In the following, we denote by $\mathcal{F}_{n,o}^T$ the set of artificial trajectories that satisfy this condition:

$$\mathcal{F}_{n,o}^T = \left\{ \left[\left(s^{l_t}, a^{l_t}, r^{l_t}, s'^{l_t} \right) \right]_{t=0}^{T-1} \in \mathcal{F}_n^T | a^{l_t} = o(t) \quad \forall t \in 0, \ldots, T-1 \right\}$$

Then, we have:

Definition (Maximal Lower Bound - Open-loop Policy o)

$$L^o(\mathcal{F}_n, s_0) = \max_{[(s^{l_t}, a^{l_t}, r^{l_t}, s'^{l_t})]_{t=0}^{T-1} \in \mathcal{F}_{n,\mathbf{o}}^T} \sum_{t=0}^{T-1} r^{l_t} - \sum_{t=0}^{T-1} L'_{Q_{T-t}} \left\| s'^{l_{t-1}} - s^{l_t} \right\|_{\mathcal{S}} .$$

Inferring safe policies from lower bounds

Similarly, a minimal upper bound $U^o(\mathcal{F}_n, s_0)$ can also be computed:

Definition (Minimal Upper Bound - Open-loop Policy$o$)

$$U^o(\mathcal{F}_n, s_0) = \min_{[(s^{l_t}, a^{l_t}, r^{l_t}, s'^{l_t})]_{t=0}^{T-1} \in \mathcal{F}_{n,o}^T} \sum_{t=0}^{T-1} r^{l_t} + \sum_{t=0}^{T-1} L'_{Q_{T-t}} \left\| s'^{l_{t-1}} - s^{l_t} \right\|_{\mathcal{S}} .$$

Both bounds are tight in the following sense:

Lemma (Tightness of the Bounds - Open-loop Policy$o$)

$$\exists C'_b > 0 : \quad V_T^o(s_0) - L^o(\mathcal{F}_n, s_0) \leq C'_b \alpha^*(\mathcal{P}_n) ,$$
$$U^o(\mathcal{F}_n, s_0) - V_T^o(s_0) \leq C'_b \alpha^*(\mathcal{P}_n) .$$

The proofs of the above stated results are given in Fonteneau et al. [2010b].

Inferring safe policies from lower bounds

We still assume that the action space $\mathcal{A}$ is finite, and we consider open-loop policies. To obtain a policy with good performance guarantees, we suggest to find an open-loop policy $\hat{o}^*_{\mathcal{F}_n,s_0} \in \mathcal{O}$ such that:

$$\hat{o}^*_{\mathcal{F}_n,s_0} \in \arg\max_{o \in \mathcal{O}} \quad L^o(\mathcal{F}_n, s_0) \ .$$

Recall that such an "open-loop" policy is optimized with respect to the initial state s_0. Solving the above optimization problem can be seen as identifying an optimal rebuilt artificial trajectory $\left[\left(s^{l_t^*}, a^{l_t^*}, r^{l_t^*}, s'^{l_t^*}\right)\right]_{t=0}^{T-1}$ and outputting as open-loop policy the sequence of actions taken along this artificial trajectory:

$$\forall t \in \{0, \ldots, T-1\}, \qquad \hat{o}^*_{\mathcal{F}_n,s_0}(t) = a^{l_t^*} \ .$$

Inferring safe policies from lower bounds

Theorem (Convergence of $\hat{o}^*_{\mathcal{F}_n,s_0}$)

Let $\mathfrak{V}^*_T(s_0)$ be the set of optimal T-step open-loop policies:

$$\mathfrak{V}^*_T(s_0) = \arg\max_{o \in \mathcal{O}} \quad V^o_T(s_0) \ ,$$

and let us suppose that $\mathfrak{V}^*_T(s_0) \neq \mathcal{O}$ (if $\mathfrak{V}^*_T(s_0) = \mathcal{O}$, the search for an optimal policy is indeed trivial). We define

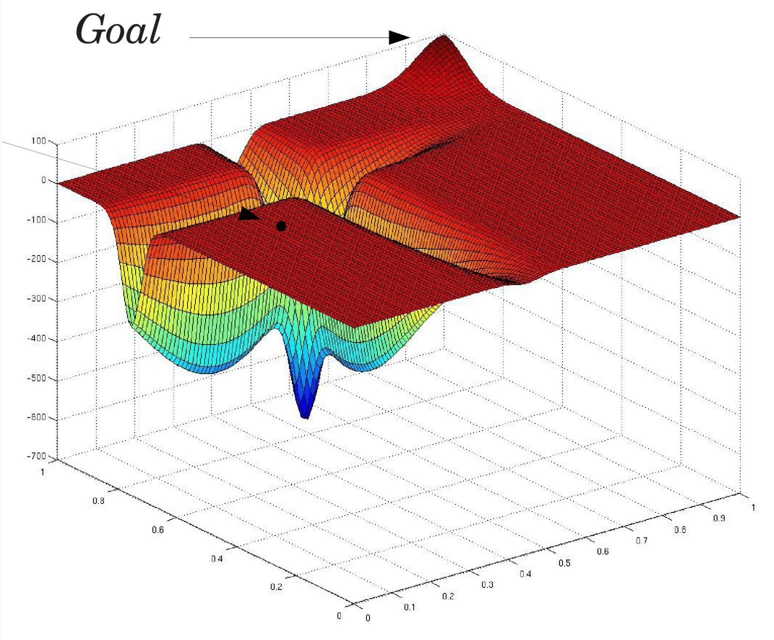
$$\varepsilon(s_0) = \min_{o \in \mathcal{O} \setminus \mathfrak{V}^*_T(s_0)} \left\{ \left(\max_{o' \in \mathcal{O}} V^{o'}_T(s_0) \right) - V^o_T(s_0) \right\} \ .$$

Then,

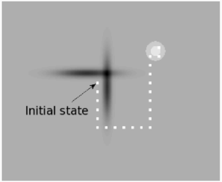
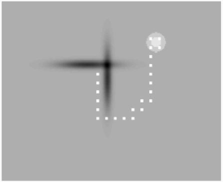
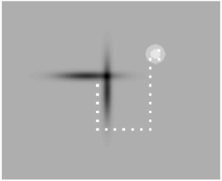
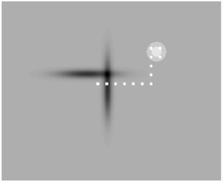
$$\left(C'_b \alpha^*(\mathcal{P}_n) < \varepsilon(s_0) \right) \implies \hat{o}^*_{\mathcal{F}_n,s_0} \in \mathfrak{V}^*_T(s_0) \ .$$

The proof of this result is also given in Fonteneau et al. [2010b].

Inferring safe policies from lower bounds



Inferring safe policies from lower bounds

	CGRL	FQI (Fitted Q Iteration)
The state space is uniformly covered by the sample		
Information about the Puddle area is removed		

Theorem (Optimal Policies computed from Optimal Trajectories)

Let $o^*_{s_0} \in \mathfrak{V}^*_T(s_0)$ *be an optimal open-loop policy. Let us assume that one can find in* $\mathcal{F}_n$ *a sequence of* T *one-step system transitions*

$$\left[\left(s^{l_0}, a^{l_0}, r^{l_0}, s^{l_1}\right), \left(s^{l_1}, a^{l_1}, r^{l_1}, s^{l_2}\right), \ldots, \left(s^{l_{T-1}}, a^{l_{T-1}}, r^{l_{T-1}}, s^{l_T}\right)\right] \in \mathcal{F}_n^T$$

such that

$$\begin{aligned} s^{l_0} &= s_0 \ , \\ a^{l_t} &= o^*_{s_0}(t) \qquad \forall t \in \{0, \ldots, T-1\} \ . \end{aligned}$$

Let $\hat{o}^*_{\mathcal{F}_n, s_0}$ *be such that*

$$\hat{o}^*_{\mathcal{F}_n, s_0} \in \underset{o \in \mathcal{O}}{\arg\max} \qquad L^o(\mathcal{F}_n, s_0) \ .$$

Then,

$$\hat{o}^*_{\mathcal{F}_n, s_0} \in \mathfrak{V}^*_T(s_0) \ .$$

Beyond the batch

- Suppose that additional system transitions can be generated.
- We detail hereafter a sampling strategy to select state-action pairs (s, a) for generating $f(s, a)$ and $r(s, a)$ so as to be able to discriminate rapidly – as new one-step transitions are generated – between optimal and non-optimal policies from $\mathcal{O}$.

First, note that a policy can only be optimal given a set of one-step transitions $\mathcal{F}$ if its upper bound is not lower than the lower bound of any element of $\mathcal{O}$. We qualify as "candidate optimal policies given $\mathcal{F}$" and we denote by $\mathcal{O}(\mathcal{F}, s_0)$ the set of policies which satisfy this property:

Definition (Candidate Optimal Policies Given $\mathcal{F}$)

$$\mathcal{O}(\mathcal{F}, s_0) = \left\{ o \in \mathcal{O} \quad \middle| \quad \forall o' \in \mathcal{O}, U^{o}(\mathcal{F}, s_0) \geq L^{o'}(\mathcal{F}, s_0) \right\}.$$

We also define the set of "compatible transitions given $\mathcal{F}$" as follows:

Definition (Compatible Transitions Given $\mathcal{F}$)

A transition $(s, a, r, s') \in \mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}$ is said compatible with the set of transitions $\mathcal{F}$ if

$$\forall (s^l, a^l, r^l, s'^l) \in \mathcal{F}, \quad \left(a^l = a\right) \implies \begin{cases} \left|r - r^l\right| & \leq \quad L_r \|s - s^l\|_{\mathcal{X}}, \\ \left\|s' - s'^l\right\|_{\mathcal{S}} & \leq \quad L_f \|s - s^l\|_{\mathcal{X}} \ . \end{cases}$$

We denote by $\mathcal{C}(\mathcal{F}) \subset \mathcal{S} \times \mathcal{U} \times \mathbb{R} \times \mathcal{U}$ the set that gathers all transitions that are compatible with the set of transitions $\mathcal{F}$.

Beyond the batch

The sampling strategy generates new one-step transitions iteratively : Given an existing set $\mathcal{F}_m$ of $m \in \mathbb{N} \setminus \{0\}$ one-step transitions, which is made of the elements of the initial set $\mathcal{F}_n$ and the m-n one-step transitions generated during the first m-n iterations of this algorithm, it selects as next sampling point $(s^{m+1}, a^{m+1}) \in \mathcal{S} \times \mathcal{A}$, the point that minimizes in the worst conditions the largest bound width among the candidate optimal policies at the next iteration:

$$(s^{m+1}, a^{m+1}) \in \underset{(s,a) \in \mathcal{S} \times \mathcal{A}}{\arg\min} \left\{ \max_{\substack{(r,s') \in \mathbb{R} \times \mathcal{S} \ s.t. (s,a,r,s') \in \mathcal{C}(\mathcal{F}_m) \\ o \in \mathcal{O}(\mathcal{F}_m \cup \{(s,a,r,s')\}, s_0)}} \delta^o \left(\mathcal{F}_m \cup \{(s,a,r,s')\}, s_0 \right) \right\}$$

where

$$\delta^o(\mathcal{F}, s_0) = U^o(\mathcal{F}, s_0) - L^o(\mathcal{F}, s_0) \; .$$

Beyond the batch

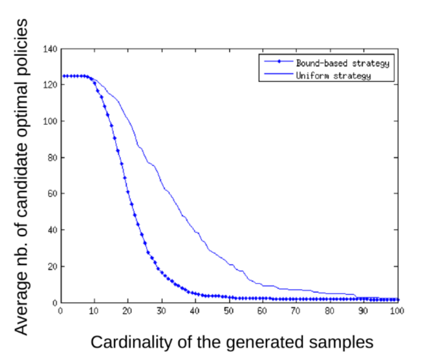
Action space : $\mathcal{A} = \{-0.20, -0.10, 0, 0.10, 0.20\}$

Dynamics and reward function : $f(s, a) = s + a$ and $r(s, a) = s + a$

Horizon : $T = 3$

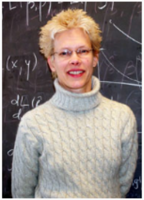
Initial state : $s_0 = -0.65$

Total number of policies : $5^3 = 125$



Many Thanks !

Many Thanks to Susan A. Murphy, Louiw Wehenkel and Damien Ernst for this collaboration.

References

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. Human-Level Control through Deep Reinforcement Learning. *Nature*, 518 (7540):529–533, 2015.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Herbert Robbins. Some aspects of the sequential design of experiments. 1952.

Rémi Munos et al. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

J Craig Nelson. The star* d study: a four-course meal that leaves us wanting more, 2006.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.

Raphael Fonteneau, Susan A Murphy, Louis Wehenkel, and Damien Ernst. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of operations research*, 208:383–416, 2013.

Raphael Fonteneau, Susan Murphy, Louis Wehenkel, and Damien Ernst. Model-free monte carlo-like policy evaluation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 217–224. JMLR Workshop and Conference Proceedings, 2010a.

Raphael Fonteneau. Contributions to batch mode reinforcement learning. 2011.

Raphael Fonteneau, Susan Murphy, Louis Wehenkel, and Damien Ernst. Inferring bounds on the performance of a control policy from a sample of trajectories. In *2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pages 117–123. IEEE, 2009.

Raphael Fonteneau, Susan Murphy, Louis Wehenkel, and Damien Ernst. A cautious approach to generalization in reinforcement learning. In *2nd International Conference on Agents and Artificial Intelligence*, 2010b.