Model-Free Monte Carlo–like Policy Evaluation

Raphael Fonteneau
University of Liège

Susan A. Murphy
University of Michigan

Louis Wehenkel
University of Liège

Damien Ernst
University of Liège

Abstract

We propose an algorithm for estimating the finite-horizon expected return of a closed loop control policy from an a priori given (off-policy) sample of one-step transitions. It averages cumulated rewards along a set of "broken trajectories" made of one-step transitions selected from the sample on the basis of the control policy. Under some Lipschitz continuity assumptions on the system dynamics, reward function and control policy, we provide bounds on the bias and variance of the estimator that depend only on the Lipschitz constants, on the number of broken trajectories used in the estimator, and on the sparsity of the sample of one-step transitions.

1 Introduction

Discrete-time stochastic optimal control problems arise in many fields such as finance, medicine, engineering as well as artificial intelligence. Many techniques for solving such problems use an oracle that evaluates the performance of any given policy in order to navigate rapidly in the space of candidate optimal policies to a (near-)optimal one.

When the considered system is accessible to experimentation at low cost, such an oracle can be based on a Monte Carlo (MC) approach. With such an approach, several "on-policy" trajectories are generated by collecting information from the system when controlled by the given policy, and the cumulated rewards observed along these trajectories are averaged to get an unbiased estimate of the performance of that policy. However if obtaining trajectories under a given policy is very costly, time consuming or otherwise difficult, e.g. in medicine or in safety critical problems, the above approach is not feasible.

In this paper, we propose a policy evaluation oracle in a *model-free* setting. In our setting, the only information available on the optimal control problem is contained in a sample of one-step transitions of the system, that have been gathered by some arbitrary experimental protocol, i.e. independently of the policy that has to be evaluated.

Our estimator is inspired by the MC approach. Similarly to the MC estimator, it evaluates the performance of a policy by averaging the sums of rewards collected along several trajectories. However, rather than "real" on-policy trajectories of the system generated by fresh experiments, it uses a set of "broken trajectories" that are rebuilt from the given sample and from the policy that is being evaluated. Under some Lipschitz continuity assumptions on the system dynamics, reward function and policy, we provide bounds on the bias and variance of our model-free policy evaluator, and show that it behaves like the standard MC estimator when the sample sparsity decreases towards zero.

The core of the paper is organized as follows. Section 2 discusses related work, Section 3 formalizes the problem, and Section 4 states our algorithm and its theoretical properties. Section 5 provides some simulation results. Proofs of our main theorems are sketched in the Appendix.

2 Related work

Model-free policy evaluation has been well studied, in particular in reinforcement learning. This field has mostly focused on the estimation of the *value function* that maps initial states into returns of the policy from these states. Temporal Difference methods (Sutton, 1988; Watkins and Dayan, 1992; Rummery and Niranjan, 1994; Bradtke and Barto, 1996) are techniques for estimating value functions from the sole knowledge of one-step transitions of the system, and their underlying theory has been well investigated, e.g., (Dayan, 1992; Tsitsiklis, 1994). In large state-spaces, these approaches have to be combined with function approximators to compactly represent the value function (Sutton et al., 2009). More recently, batch-mode approximate value iteration algorithms have been successful in using function approximators to estimate value functions in a model-free setting (Ormoneit and Sen, 2002; Ernst et al., 2005; Riedmiller, 2005), and several papers have analyzed some of their theoretical properties (Antos et al., 2007; Munos and Szepesvári, 2008).

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

The Achilles' heel of all these techniques is their strong dependence on the choice of a suitable function approximator, which is not straightforward (Busoniu et al., 2010). Contrary to these techniques, the estimator proposed in this paper does not use function approximators. As mentioned above, it is an extension of the standard MC estimator to a model-free setting, and in this, it is related to current work seeking to build computationally efficient model-based Monte Carlo estimators, e.g., (Dimitrakakis and Lagoudakis, 2008).

3 Problem statement

We consider a discrete-time system whose behavior over T stages is characterized by a time-invariant dynamics $x_{t+1} = f(x_t, u_t, w_t) \quad t = 0, 1, \ldots, T-1$, where x_t belongs to a normed vector space $\mathcal{X}$ of states, and u_t belongs to a normed vector space $\mathcal{U}$ of control actions. An instantaneous reward $r_t = \rho(x_t, u_t, w_t) \in \mathbb{R}$ is associated with the transition from t to $t+1$. The stochasticity of the control problem is induced by the unobservable random process $w_t \in \mathcal{W}$, which we suppose to be drawn i.i.d. according to a probability distribution $p_{\mathcal{W}}(.)$, $\forall t = 0, \ldots, T-1$. In the following, we signal this by $w_t \sim p_{\mathcal{W}}(.)$ and, as induced by the notation, we assume that $p_{\mathcal{W}}(.)$ depends neither on (x_t, u_t) nor on $t \in [\![0, T-1]\!]$ (using the notation $[\![0, T-1]\!] = \{0, \ldots, T-1\}$) . $T \in \mathbb{N}_0$ is referred to as the optimization horizon of the control problem.

Let $h : [\![0, T-1]\!] \times \mathcal{X} \to \mathcal{U}$ be a deterministic closed-loop time-varying control policy that maps the time t and the current state x_t into the action $u_t = h(t, x_t)$, and let $J^h(x_0)$ denote the expected return of this policy h, defined as follows :

$$J^h(x_0) = \mathop{\mathbb{E}}_{w_0, \ldots, w_{T-1} \sim p_{\mathcal{W}}(.)} \left[R^h(x_0) \right] ,$$

where $R^h(x_0) = \sum_{t=0}^{T-1} \rho(x_t, h(t, x_t), w_t)$ and $x_{t+1} = f(x_t, h(t, x_t), w_t)$. A realization of the random variable $R^h(x_0)$ corresponds to the cumulated reward of h when used to control the system from the initial condition x_0 over T stages while disturbed by the random process $w_t \sim p_{\mathcal{W}}(.)$. We suppose that $R^h(x_0)$ has a finite variance $\sigma^2_{R^h}(x_0) = \mathop{Var}_{w_0, \ldots, w_{T-1} \sim p_{\mathcal{W}}(.)} \left[R^h(x_0) \right]$.

In our setting, f, ρ and $p_{\mathcal{W}}(.)$ are fixed but *unknown* (and hence inaccessible to simulation). The only information available on the control problem is gathered in a given sample of n one-step transitions $\mathcal{F}_n = [(x^l, u^l, r^l, y^l)]_{l=1}^n$, where the first two elements (x^l and u^l) of every one-step transition are chosen in an arbitrary way, while the pairs (r^l, y^l) are consistently determined by $(\rho(x^l, u^l, .), f(x^l, u^l, .))$, drawn according to $p_{\mathcal{W}}(.)$. We want to estimate from such a sample $\mathcal{F}_n$, the expected return $J^h(x_0)$ of the given policy h for a given initial state x_0.

4 A model-free Monte Carlo–like estimator of $J^h(x_0)$

We first remind the classical model-based MC estimator and its bias and variance in Section 4.1. In Section 4.2 we explain our estimator which mimics the MC estimator in a model-free setting, and in Section 4.3 we provide a theoretical analysis of the bias and variance of this estimator.

4.1 Model-based MC estimator

The MC estimator works in a model-based setting (i.e., in a setting where f, ρ and $p_{\mathcal{W}}(.)$ are known). It estimates $J^h(x_0)$ by averaging the returns of several (say $p \in \mathbb{N}_0$) trajectories of the system which have been generated by simulating the system from x_0 using the policy h. More formally, the MC estimator of the expected return of the policy h when starting from the initial state x_0 writes

$$\mathbb{M}^h_p(x_0) = \frac{1}{p} \sum_{i=1}^{p} \sum_{t=0}^{T-1} \rho(x^i_t, h(t, x^i_t), w^i_t)$$

with $\forall t \in [\![0, T-1]\!], \forall i \in [\![1, p]\!]$: $w^i_t \sim p_{\mathcal{W}}(.), x^i_0 = x_0$, $x^i_{t+1} = f(x^i_t, h(t, x^i_t), w^i_t)$. It is well known that the bias and variance of the MC estimator are

$$\mathop{\mathbb{E}}_{w^i_t \sim p_{\mathcal{W}}(.), i=1\ldots p, t=0\ldots T-1} \left[\mathbb{M}^h_p(x_0) - J^h(x_0) \right] = 0 ,$$

$$\mathop{Var}_{w^i_t \sim p_{\mathcal{W}}(.), i=1\ldots p, t=0\ldots T-1} \left[\mathbb{M}^h_p(x_0) \right] = \frac{\sigma^2_{R^h}(x_0)}{p} .$$

4.2 Model-free MC estimator

From a sample $\mathcal{F}_n$, our model-free MC (MFMC) estimator works by selecting p sequences of transitions of length T from this sample that we call "broken trajectories". These broken trajectories will then serve as proxies of p "actual" trajectories that could be obtained by simulating the policy h on the given control problem. Our estimator averages the cumulated returns over these broken trajectories to compute its estimate of $J^h(x_0)$. The main idea behind our method consists of selecting the broken trajectories so as to minimize the discrepancy of these trajectories with a classical MC sample that could be obtained by simulating the system with policy h.

To build a sample of p substitute broken trajectories of length T starting from x_0 and similar to trajectories that would be induced by a policy h, our algorithm uses each one-step transition in $\mathcal{F}_n$ at most once; we thus assume that $pT \leq n$. The p broken trajectories of T one-step transitions are created sequentially. Every broken trajectory is grown in length by selecting, among the sample of not yet used one-step transitions, a transition whose first two elements minimize the distance $-$ using a distance metric Δ in $\mathcal{X} \times \mathcal{U}$ $-$ with the couple formed by the last element of

```
MFMC sampling (arguments: $\mathcal{F}_n, h(.,.), x_0, \Delta(.,.), T, p$)

    Let $\mathcal{G}$ denote the current set of not yet used one-step
    transitions in $\mathcal{F}_n$; initially, set $\mathcal{G} = \mathcal{F}_n$;

    For $i = 1$ to $p$, extract a broken trajectory by doing:

        Set $t = 0$ and $x_t^i = x_0$;
        While $t < T$ do
            Set $u_t^i = h(t, x_t^i)$; then compute the set
            $\mathcal{H} = \underset{(x,u,r,y) \in \mathcal{G}}{\arg\min} \left(\Delta((x,u),(x_t^i, u_t^i))\right)$;
            Let $l_t^i$ be the lowest index in $\mathcal{F}_n$ of the transitions that belong to $\mathcal{H}$;
            Set $t = t+1$, $x_t^i = y^{l_t^i}$;
            Set $\mathcal{G} = \mathcal{G} \setminus \{(x^{l_t^i}, u^{l_t^i}, r^{l_t^i}, y^{l_t^i})\}$;
        end While

    end For

    Return the set of indices $\{l_t^i\}_{i=1,t=0}^{i=p,t=T-1}$.
```

Figure 1: MFMC algorithm to generate a set of size p of $T-$length broken trajectories from a sample of n one-step transitions.

the previously selected transition and the action induced by h at the end of this previous transition.

A tabular version of the algorithm for building the broken trajectories is given on Figure 1. It returns a set of indices of one-step transitions $\{l_t^i\}_{i=1,t=0}^{i=p,t=T-1}$ from $\mathcal{F}_n$ based on h, x_0, the distance metric Δ and the parameter p. Based on this set of indices, we define our MFMC estimate of the expected return of the policy h when starting from the initial state x_0 by:

$$\mathfrak{M}_p^h(\mathcal{F}_n, x_0) \quad = \quad \frac{1}{p}\sum_{i=1}^{p}\sum_{t=0}^{T-1} r^{l_t^i}.$$

Figure 2 illustrates the MFMC estimator. Note that the computation of the MFMC estimator $\mathfrak{M}_p^h(\mathcal{F}_n, x_0)$ has a linear complexity with respect to the cardinality n of $\mathcal{F}_n$ and the length T of the broken trajectories.

4.3 Analysis of the MFMC estimator

In this section we characterize some main properties of our estimator. To this end, we proceed as follows:

1. we first abstract away from the given sample $\mathcal{F}_n$ by instead considering an ensemble of samples of pairs which are "compatible" with $\mathcal{F}_n$ in the following sense: from $\mathcal{F}_n = [(x^l, u^l, r^l, y^l)]_{l=1}^n$, we keep only the sample $\mathcal{P}_n = [(x^l, u^l)]_{l=1}^n \in (\mathcal{X} \times \mathcal{U})^n$ of state-action pairs, and we then consider the ensemble of samples of one-step transitions of size n that could be generated by completing each pair (x^l, u^l) of $\mathcal{P}_n$ by drawing for each l a disturbance signal w^l at random from $p_{\mathcal{W}}(.)$, and by recording the resulting values of $f(x^l, u^l, w^l)$ and $\rho(x^l, u^l, w^l)$. We denote by $\tilde{\mathcal{F}}_n$ one such "random" set of one-step transitions defined by a random draw of n disturbance signals $w^l \quad l = 1 \ldots n$. The sample of one-step transitions $\mathcal{F}_n$ is thus a realization of the random set $\tilde{\mathcal{F}}_n$;

2. we then study the distribution of our estimator $\mathfrak{M}_p^h(\tilde{\mathcal{F}}_n, x_0)$, seen as a function of the random set $\tilde{\mathcal{F}}_n$; in order to characterize this distribution, we express its bias and its variance as a function of a measure of the density of the sample $\mathcal{P}_n$, defined by its "$k-$sparsity"; this is the smallest radius such that all Δ-balls in $\mathcal{X} \times \mathcal{U}$ of this radius contain at least k elements from $\mathcal{P}_n$. The use of this notion implies that the space $\mathcal{X} \times \mathcal{U}$ is bounded (when measured using the distance metric Δ).

The bias and variance characterization will be done under some additional assumptions detailed below. After that, we state the main theorems formulating these characterizations. Proofs are given in the Appendix.

Lipschitz continuity of the functions f, ρ and h. We assume that the dynamics f, the reward function ρ and the policy h are Lipschitz continuous, i.e., we assume that the states and actions belong to a normed vector space and that there exist finite constants L_f, L_ρ and $L_h \in \mathbb{R}^+$ such that $\forall\, (x, x', u, u', w) \in \mathcal{X}^2 \times \mathcal{U}^2 \times \mathcal{W}$,
$\|f(x,u,w) - f(x',u',w)\|_{\mathcal{X}} \le L_f(\|x - x'\|_{\mathcal{X}} + \|u - u'\|_{\mathcal{U}})$,
$|\rho(x,u,w) - \rho(x',u',w)| \le L_\rho(\|x - x'\|_{\mathcal{X}} + \|u - u'\|_{\mathcal{U}})$,
$\forall t \in [\![0, T-1]\!], \|h(t,x) - h(t,x')\|_{\mathcal{U}} \le L_h \|x - x'\|_{\mathcal{X}}$,
where $\|.\|_{\mathcal{X}}$ and $\|.\|_{\mathcal{U}}$ denote the chosen norms over the spaces $\mathcal{X}$ and $\mathcal{U}$, respectively.

Distance metric Δ and $k-$sparsity of a sample $\mathcal{P}_n$. We assume that $\forall (x, x', u, u') \in \mathcal{X}^2 \times \mathcal{U}^2, \Delta((x,u),(x',u')) = (\|x - x'\|_{\mathcal{X}} + \|u - u'\|_{\mathcal{U}})$. We suppose that $\mathcal{X} \times \mathcal{U}$ is bounded when measured using the distance metric Δ, and, given $k \in \mathbb{N}_0$ with $k \le n$, we define the *$k-$sparsity*, $\alpha_k(\mathcal{P}_n)$ of the sample $\mathcal{P}_n$ by $\alpha_k(\mathcal{P}_n) = \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} \left\{\Delta_k^{\mathcal{P}_n}(x,u)\right\}$, where $\Delta_k^{\mathcal{P}_n}(x,u)$ denotes the distance of (x,u) to its $k-$th nearest neighbor (using the distance metric Δ) in the $\mathcal{P}_n$ sample.

Bias of the MFMC estimator. We propose to compute an upper bound of the bias and variance of the MFMC estimator. To this end, we denote by $E_{p,\mathcal{P}_n}^h(x_0)$ the expected value:

$$E_{p,\mathcal{P}_n}^h(x_0) = \underset{w^1,\ldots,w^n \sim p_{\mathcal{W}}(.)}{\mathbb{E}}\left[\mathfrak{M}_p^h(\tilde{\mathcal{F}}_n, x_0)\right].$$

We have the following theorem (proof in Appendix A):

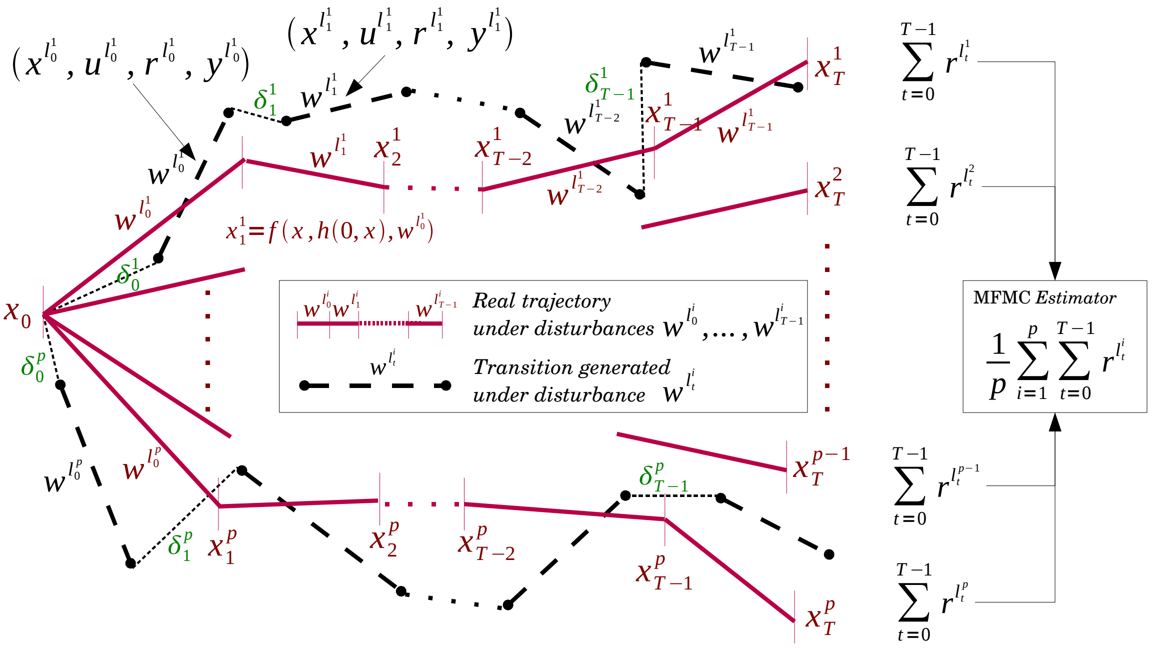
$(x^{l_0^1}, u^{l_0^1}, r^{l_0^1}, y^{l_0^1})$
$(x^{l_1^1}, u^{l_1^1}, r^{l_1^1}, y^{l_1^1})$
x_0
δ_0^1
δ_1^1
δ_{T-1}^1
$w^{l_0^1}$
$w^{l_1^1}$
$w^{l_{T-2}^1}$
$w^{l_{T-1}^1}$
x_2^1
x_{T-2}^1
x_{T-1}^1
x_T^1
x_T^2
x_T^{p-1}
$x_1^1=f(x,h(0,x),w^{l_0^1})$
δ_0^p
δ_1^p
δ_{T-1}^p
$w^{l_0^p}$
x_1^p
x_2^p
x_{T-2}^p
x_{T-1}^p
x_T^p
$w^{l_0^i}\ w^{l_1^i}\ w^{l_{T-1}^i}$ Real trajectory under disturbances $w^{l_0^i},\dots,w^{l_{T-1}^i}$
$w^{l_t^i}$ Transition generated under disturbance $w^{l_t^i}$
$\sum_{t=0}^{T-1} r^{l_t^1}$
$\sum_{t=0}^{T-1} r^{l_t^2}$
$\sum_{t=0}^{T-1} r^{l_t^{p-1}}$
$\sum_{t=0}^{T-1} r^{l_t^p}$
MFMC Estimator
$\frac{1}{p}\sum_{i=1}^{p}\sum_{t=0}^{T-1} r^{l_t^i}$

Figure 2: The MFMC estimator builds p broken trajectories made of one-step transitions.

Theorem 4.1 (Bias of the MFMC estimator)

$$\left|J^h(x_0) - E^h_{p,\mathcal{P}_n}(x_0)\right| \le C\alpha_{pT}(\mathcal{P}_n)$$

$$\textit{with } C = L_\rho \sum_{t=0}^{T-1}\sum_{i=0}^{T-t-1}\left[L_f(1+L_h)\right]^i .$$

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 $V^h_{p,\mathcal{P}_n}(x_0)$ the variance of the MFMC estimator defined by

$$V^h_{p,\mathcal{P}_n}(x_0) = \underset{w^1,\dots,w^n\sim p_{\mathcal{W}}(.)}{Var}\left[\mathfrak{M}^h_p(\tilde{\mathcal{F}}_n,x_0)\right]$$

$$= \underset{w^1,\dots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\left[\left(\mathfrak{M}^h_p(\tilde{\mathcal{F}}_n,x_0) - E^h_{p,\mathcal{P}_n}(x_0)\right)^2\right]$$

and we give the following theorem.

Theorem 4.2 (Variance of the MFMC estimator)

$$V^h_{p,\mathcal{P}_n}(x_0) \le \left(\frac{\sigma_{R^h}(x_0)}{\sqrt{p}} + 2C\alpha_{pT}(\mathcal{P}_n)\right)^2$$

$$\textit{with } C = L_\rho \sum_{t=0}^{T-1}\sum_{i=0}^{T-t-1}\left[L_f(1+L_h)\right]^i .$$

The proof of this theorem is given in Appendix B. We see that the variance of our MFMC estimator is guaranteed to be close to that of the classical MC estimator if the sample sparsity is small enough. Note, however, that our bounds are quite conservative given the very weak assumptions that we exploit about the considered optimal control problem.

5 Illustration

In this section, we illustrate the MFMC estimator on an academic problem.

Problem statement. The system dynamics and the reward function are given by $x_{t+1} = \sin\left(\frac{\pi}{2}(x_t + u_t + w_t)\right)$ and $\rho(x_t,u_t,w_t) = \frac{1}{2\pi}e^{-\frac{1}{2}(x_t^2+u_t^2)} + w_t$ with the state space $\mathcal{X}$ being equal to $[-1,1]$ and the action space $\mathcal{U}$ to $[-\frac{1}{2},\frac{1}{2}]$. The disturbance w_t is an element of the interval $\mathcal{W} = [-\frac{\epsilon}{2},\frac{\epsilon}{2}]$ with $\epsilon = 0.1$ and $p_{\mathcal{W}}$ is a uniform probability distribution over the interval $\mathcal{W}$. The optimization horizon T is equal to 15. The policy h whose performances have to be evaluated writes $h(t,x) = -\frac{x}{2}, \forall x \in \mathcal{X}, \forall t \in [\![0,T-1]\!]$. The initial state of the system is set $x_0 = -0.5$. The samples of one-step transitions $\mathcal{F}_n$ that are used as substitute for f, ρ and $p_{\mathcal{W}}(.)$ in our experiments have been generated according to the mechanism described in Section 4.3.

Results. For our first set of experiments, we choose to work with a value of $p = 10$ i.e., the MFMC estimator rebuilds 10 broken trajectories to estimate $J^h(-0.5)$. In these experiments, for different cardinalities $n_j = (10j)^2 \quad j = 1\dots 10$, we generate 50 sets $\mathcal{F}^1_{n_j},\dots,\mathcal{F}^{50}_{n_j}$

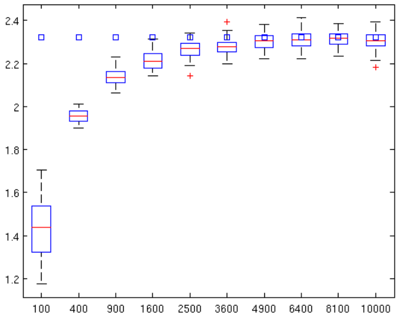

Figure 3: Computations of the MFMC estimator for different cardinalities of the sample of one-step transitions with $p = 10$. Squares represent $J^h(x_0)$.

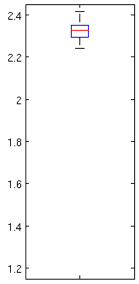

Figure 4: Computations of the MC estimator with $p = 10$.

and run our MFMC estimator on each of these sets. For a given cardinality $n_j = m_j^2$, all the different samples $\mathcal{F}^1_{n_j}, \ldots, \mathcal{F}^{50}_{n_j}$ are generated considering the same couples $(x^l, u^l) \quad l = 1 \ldots n_j$ that uniformly cover the space according to the relationships $x^l = -1 + \frac{2j_1}{m_j}$ and $u^l = -1 + \frac{2j_2}{m_j}$ with $j_1, j_2 \in [\![0, m_j - 1]\!]$. The results of this first set of experiments are gathered in Figure 3. For every value of n_j considered in our experiments, the 50 values outputted by the MFMC estimator are concisely represented by a box plot. The box has lines at the lower quartile, median, and upper quartile values. Whiskers extend from each end of the box to the adjacent values in the data within 1.5 times the interquartile range from the ends of the box. Outliers are data with values beyond the ends of the whiskers and are displayed with a red $+$ sign. The squares represent an accurate estimate of $J^h(-0.5)$ computed by running thousands of Monte Carlo simulations. As we observe, when the samples increase in size (which corresponds to a decrease of the $pT-$sparsity $\alpha_{pT}(\mathcal{P}_n)$) the MFMC estimator is more likely to output accurate estimations of $J^h(-0.5)$. As explained throughout this paper, there exist many similarities between the model-free MFMC estimator and the model-based MC estimator. These can be empirically illustrated by putting Figure 3 in perspective with Figure 4. This figure reports the results obtained by 50 independent runs of the MC estimator, every of these runs using also $p = 10$ trajectories. As expected, one can see that the MFMC estimator tends to behave similarly to the MC estimator when the cardinality of the sample increases.

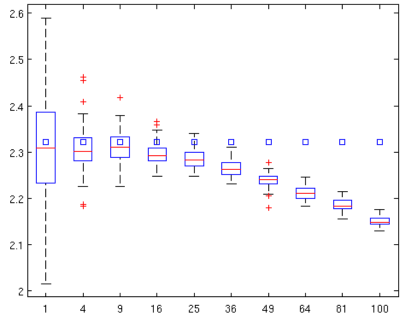

Figure 5: Computations of the MFMC estimator for different values of the number of broken trajectories p. Squares represent $J^h(x_0)$.

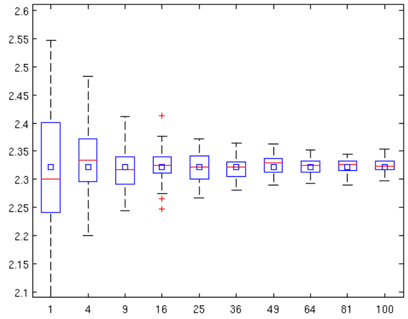

Figure 6: Computations of the MC estimator for different values of the number of broken trajectories p. Squares represent $J^h(x_0)$.

In our second set of experiments, we choose to study the influence of the number of broken trajectories p upon which the MFMC estimator bases its prediction. In these experiments, for each value $p_j = j^2 \quad j = 1 \ldots 10$ we generate 50 samples $\mathcal{F}^1_{10,000}, \ldots, \mathcal{F}^{50}_{10,000}$ of one-step transitions of cardinality $10,000$ and use these samples to compute the MFMC estimator. The results are plotted in Figure 5. This figure shows that the bias of the MFMC estimator seems to be relatively small for small values of p and to increase with p. This is in accordance with Theorem 4.1 which bounds the bias with an expression that is increasing with p.

In Figure 6, we have plotted the evolution of the values outputted by the model-based MC estimator when the number of trajectories it considers in its prediction increases. While, for small number of trajectories, it behaves similarly to the MFMC estimator, the quality of its predictions steadily increases with p, while it is not the case for the MFMC estimator whose performances degrade once p crosses a threshold value. Notice that this threshold value could be made larger by increasing the size of the samples of one-step system transitions used as input of the MFMC algorithm.

6 Conclusions

We have proposed in this paper an estimator of the expected return of a policy in a model-free setting. The estimator named MFMC works by rebuilding from a sample of one-step transitions a set of broken trajectories and by averaging the sum of rewards gathered along these latter trajectories. In this respect, it can be seen as an extension to a model-free setting of the standard model-based Monte Carlo policy evaluation technique. We have provided bounds on the bias and variance of the MFMC estimator ; these were depending among others on the sparsity of the sample of one-step transitions and the Lipschitz constants associated with the system dynamics, reward function and policy. These bounds show that when the sample sparsity becomes small, the bias of the estimator decreases to zero and its variance converges to the variance of the Monte Carlo estimator.

The work presented in this paper could be extended along several lines. For example, it would be interesting to consider disturbances whose probability distributions are conditioned on the states and the actions and to study how the bounds given in this paper should be modified to remain valid in such a setting. Another interesting research direction would be to investigate how the bounds proposed in this paper could be useful for choosing automatically the parameters of the MFMC estimator which are the number p of broken trajectories it rebuilds and the distance metric Δ it uses to select its set of broken trajectories.

However, the bound on the variance of the MFMC estimator depends explicitly on the “natural” variance of the sum of rewards along trajectories of the system when starting from the same initial state. Using this bound for determining automatically p (and/or Δ) suggests therefore to investigate how an upper bound on this natural variance could be inferred from the sample of one-step transitions. Finally, this MFMC estimator adds to the arsenal of techniques that have been proposed in the literature for computing an estimate of the expected return of a policy in a model-free setting. However, it is not yet clear how it would compete with such techniques. All these techniques have pros and cons and establishing which one to exploit for a specific problem certainly deserves further research.

Acknowledgements

Raphael Fonteneau acknowledges the financial support of the FRIA. Damien Ernst is a research associate of the FRS-FNRS. This paper presents research results of the Belgian Network BIOMAGNET and the PASCAL2 European Network of Excellence. We also acknowledge financial support from NIH grants P50 DA10075 and R01 MH080015. The scientific responsibility rests with its authors.

References

A. Antos, R. Munos, and C. Szepesvári. Fitted Q-iteration in continuous action space MDPs. In *Advances in Neural Information Processing Systems 20, NIPS 2007*, 2007.

S.J. Bradtke and A.G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57, 1996.

L. Busoniu, R. Babuska, B. De Schutter, and D. Ernst. *Reinforcement Learning and Dynamic Programming using Function Approximators*. Taylor & Francis CRC Press, 2010.

P. Dayan. The convergence of TD(λ) for general λ. *Machine Learning*, 8:341–162, 1992.

C Dimitrakakis and M. G. Lagoudakis. Rollout sampling approximate policy iteration. *Machine Learning*, 72: 157–171, 2008.

D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, pages 815–857, 2008.

D. Ormoneit and S. Sen. Kernel-based reinforcement learning. *Machine Learning*, 49(2-3):161–178, 2002.

M. Riedmiller. Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method. In *Proceedings of the Sixteenth European Conference on Machine Learning (ECML 2005)*, pages 317–328, 2005.

G.A. Rummery and M. Niranjan. On-line Q-learning using connectionist systems. Technical Report 166, Cambridge University Engineering Department, 1994.

R.S. Sutton. Learning to predict by the methods of temporal difference. *Machine Learning*, 3:9–44, 1988.

R.S. Sutton, H. Reza Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

J.N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.

C.J. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):179–192, 1992.

Appendix

A Proof of Theorem 4.1

Before giving the proof of Theorem 4.1, we first give three preliminary lemmas. Given a disturbance vector $\Omega = [\Omega(0), \ldots, \Omega(T-1)] \in \mathcal{W}^T$, we define the Ω-disturbed state-action value function $Q^{h,\Omega}_{T-t}(x,u)$ for $t \in \llbracket 0, T-1 \rrbracket$ as follows: $Q^{h,\Omega}_{T-t}(x,u) = \rho(x,u,\Omega(t)) + \sum_{t'=t+1}^{T-1} \rho(x_{t'}, h(t', x_{t'}), \Omega(t'))$with $x_{t+1} = f(x,u,\Omega(t))$ and $x_{t'+1} = f(x_{t'}, h(t', x_{t'}), \Omega(t')), \forall t' \in \llbracket t+1, T-1 \rrbracket$. Then, we define the expected return given Ω the quantity $\mathbb{E}[R^h(x_0)|\Omega] = \underset{w_0,\ldots,w_{T-1} \sim p_{\mathcal{W}}(.)}{\mathbb{E}} [R^h(x_0)|w_0 = \Omega(0), \ldots, w_{T-1} = \Omega(T-1)]$. From there, we have the two following trivial results: $\forall (\Omega, x_0) \in \mathcal{W}^T \times \mathcal{X}$,

$$\mathbb{E}[R^h(x_0)|\Omega] = Q^{h,\Omega}_T(x_0, h(0,x_0)) \tag{1}$$

and $\forall (x,u) \in \mathcal{X} \times \mathcal{U}, \forall \Omega \in \mathcal{W}^T$,

$$Q^{h,\Omega}_{T-t+1}(x,u) = \rho(x,u,\Omega(t-1)) + Q^{h,\Omega}_{T-t}\big(f\big(x,u,\Omega(t-1)\big), h\big(t, f(x,u,\Omega(t-1))\big)\big) \ , \tag{2}$$

Then, we have the following lemma.

Lemma A.1 (Lipschitz Continuity of $Q^{h,\Omega}_{T-t}$)
$\forall t \in \llbracket 0, T-1 \rrbracket, \forall (x,x',u,u') \in \mathcal{X}^2 \times \mathcal{U}^2$,
$\big|Q^{h,\Omega}_{T-t}(x,u) - Q^{h,\Omega}_{T-t}(x',u')\big| \leq L_{Q_{T-t}} \Delta((x,u),(x',u'))$
with $L_{Q_{T-t}} = L_\rho \sum_{i=0}^{T-t-1} \big[L_f(1+L_h)\big]^i$.

Proof of Lemma A.1 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 $\big|Q^{h,\Omega}_{T-t}(x,u) - Q^{h,\Omega}_{T-t}(x',u')\big|$ by Δ^Q_{T-t}.
Basis: $t = T-1$ We have $\Delta^Q_1 = |\rho(x,u,\Omega(T-1)) - \rho(x',u',\Omega(T-1)|$, and the Lipschitz continuity of ρ allows to write $\Delta^Q_1 \leq L_\rho\big(\|x-x'\|_{\mathcal{X}} + \|u-u'\|_{\mathcal{U}}\big) = L_\rho \Delta((x,u),(x',u'))$. This proves $\mathcal{H}(1)$.
Induction step: We suppose that $\mathcal{H}(T-t)$ is true, $1 \leq t \leq T-1$. Using Equation (2), one has
$\Delta^Q_{T-t+1} = \big|Q^{h,\Omega}_{T-t+1}(x,u) - Q^{h,\Omega}_{T-t+1}(x',u')\big| = \big|\rho(x,u,\Omega(t-1)) - \rho(x',u',\Omega(t-1)) + Q^{h,\Omega}_{T-t}(f(x,u,\Omega(t-1)), h(t, f(x,u,\Omega(t-1)))) - Q^{h,\Omega}_{T-t}(f(x',u',\Omega(t-1)), h(t, f(x',u',\Omega(t-1))))\big|$ and, from there,
$\Delta^Q_{T-t+1} \leq \big|\rho(x,u,\Omega(t-1)) - \rho(x',u',\Omega(t-1))\big| + \big|Q^{h,\Omega}_{T-t}(f(x,u,\Omega(t-1)), h(t, f(x,u,\Omega(t-1)))) - Q^{h,\Omega}_{T-t}(f(x',u',\Omega(t-1)), h(t, f(x',u',\Omega(t-1))))\big|$.

$\mathcal{H}(T-t)$ and the Lipschitz continuity of ρ give
$\Delta^Q_{T-t+1} \leq L_\rho \Delta((x,u),(x',u')) + L_{Q_{T-t}} \Delta((f(x,u,\Omega(t-1)), h(t, f(x,u,\Omega(t-1)))), (f(x',u',\Omega(t-1)), h(t, f(x',u',\Omega(t-1)))))$.

Using the Lipschitz continuity of f and h, we have
$\Delta^Q_{T-t+1} \leq L_\rho \Delta((x,u),(x',u')) + L_{Q_{T-t}}\big(L_f \Delta((x,u),(x',u')) + L_h L_f \Delta((x,u),(x',u'))\big)$, and, from there, $\Delta^Q_{T-t+1} \leq L_{Q_{T-t+1}} \Delta((x,u),(x',u'))$ since $L_{Q_{T-t+1}} \doteq L_\rho + L_{Q_{T-t}} L_f (1+L_h)$. This proves $\mathcal{H}\ (T-t+1)$ and ends the proof.

Given a broken trajectory $\tau^i = [(x^{l^i_t}, u^{l^i_t}, r^{l^i_t}, y^{l^i_t})]_{t=0}^{T-1}$ we denote by Ω^{τ^i} its associated disturbance vector $\Omega^{\tau^i} = [w^{l^i_0}, \ldots, w^{l^i_{T-1}}]$, i.e. the vector made of the T unknown disturbances that affected the generation of the one-step transitions $(x^{l^i_t}, u^{l^i_t}, r^{l^i_t}, y^{l^i_t})$ (cf. first item of Section 4.3). We give the following lemma.

Lemma A.2 (Bounds on the expected return given Ω)
$\forall i \in \llbracket 1, p \rrbracket,\ b^h(\tau^i, x_0) \leq \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}] \leq a^h(\tau^i, x_0)$,
with
$b^h(\tau^i, x_0) = \sum_{t=0}^{T-1} \big[r^{l^i_t} - L_{Q_{T-t}} \delta^i_t\big]$,
$a^h(\tau^i, x_0) = \sum_{t=0}^{T-1} \big[r^{l^i_t} + L_{Q_{T-t}} \delta^i_t\big]$,
$\delta^i_t = \Delta((x^{l^i_t}, u^{l^i_t}), (y^{l^i_{t-1}}, h(t, y^{l^i_{t-1}})))$, $\forall t \in \llbracket 0, T-1 \rrbracket$,
$y^{l^i_{-1}} = x_0, \forall i \in \llbracket 1, p \rrbracket$.

Proof of Lemma A.2 Let us first prove the lower bound. With $u_0 = h(0,x_0)$, the Lipschitz continuity of $Q^{h,\Omega^{\tau^i}}_T$ gives $|Q^{h,\Omega^{\tau^i}}_T(x_0,u_0) - Q^{h,\Omega^{\tau^i}}_T(x^{l^i_0}, u^{l^i_0})| \leq L_{Q_T} \Delta((x_0,u_0),(x^{l^i_0}, u^{l^i_0}))$.
Equation (1) gives $Q^{h,\Omega^{\tau^i}}_T(x_0,u_0) = \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]$.
Thus, $\big|\mathbb{E}[R^h(x_0)|\Omega^{\tau^i}] - Q^{h,\Omega^{\tau^i}}_T(x^{l^i_0}, u^{l^i_0})\big| = \big|Q^{h,\Omega^{\tau^i}}_T(x_0, h(0,x_0)) - Q^{h,\Omega^{\tau^i}}_T(x^{l^i_0}, u^{l^i_0})\big| \leq L_{Q_T} \Delta((x_0, h(0,x_0)), (x^{l^i_0}, u^{l^i_0}))$.
It follows that
$Q^{h,\Omega^{\tau^i}}_T(x^{l^i_0}, u^{l^i_0}) - L_{Q_T} \delta^i_0 \leq \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]$. Using Equation (2) we have
$Q^{h,\Omega^{\tau^i}}_T(x^{l^i_0}, u^{l^i_0}) = \rho(x^{l^i_0}, u^{l^i_0}, w^{l^i_0}) + Q^{h,\Omega^{\tau^i}}_{T-1}\big(f(x^{l^i_0}, u^{l^i_0}, w^{l^i_0}), h(1, f(x^{l^i_0}, u^{l^i_0}, w^{l^i_0}))\big)$.
By definition of Ω^{τ^i}, we have: $\rho(x^{l^i_0}, u^{l^i_0}, w^{l^i_0}) = r^{l^i_0}$ and $f(x^{l^i_0}, u^{l^i_0}, w^{l^i_0}) = y^{l^i_0}$. From there
$Q^{h,\Omega^{\tau^i}}_T(x^{l^i_0}, u^{l^i_0}) = r^{l^i_0} + Q^{h,\Omega^{\tau^i}}_{T-1}(y^{l^i_0}, h(1, y^{l^i_0}))$,
and
$Q^{h,\Omega^{\tau^i}}_{T-1}(y^{l^i_0}, h(1, y^{l^i_0})) + r^{l^i_0} - L_{Q_T} \delta^i_0 \leq \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]$.
The Lipschitz continuity of $Q^{h,\Omega^{\tau^i}}_{T-1}$ gives
$\big|Q^{h,\Omega^{\tau^i}}_{T-1}(y^{l^i_0}, h(1, y^{l^i_0})) - Q^{h,\Omega^{\tau^i}}_{T-1}(x^{l^i_1}, u^{l^i_1})\big| \leq L_{Q_{T-1}} \Delta((y^{l^i_0}, h(1, y^{l^i_0})), (x^{l^i_1}, u^{l^i_1})) = L_{Q_{T-1}} \delta^i_1$,
which implies that
$Q^{h,\Omega^{\tau^i}}_{T-1}(x^{l^i_1}, u^{l^i_1}) - L_{Q_{T-1}} \delta^i_1 \leq Q^{h,\Omega^{\tau^i}}_{T-1}(y^{l^i_0}, h(1, y^{l^i_0}))$.
We therefore have
$Q^{h,\Omega^{\tau^i}}_{T-1}(x^{l^i_1}, u^{l^i_1}) + r^{l^i_0} - L_{Q_T} \delta^i_0 - L_{Q_{T-1}} \delta^i_1 \leq \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]$.
The proof is completed by iterating this derivation. The upper bound is proved similarly. We give a third lemma.

Lemma A.3 $\forall i \in [\![1,p]\!], a^h(\tau^i,x_0) - b^h(\tau^i,x_0) \leq 2C\alpha_{pT}(\mathcal{P}_n)$ *with* $C = \sum_{t=0}^{T-1} L_{Q_{T-t}}$.

Proof of Lemma A.3 By construction of the bounds, one has $a^h(\tau^i,x_0) - b^h(\tau^i,x_0) = \sum_{t=0}^{T-1} 2L_{Q_{T-t}}\delta_t^i$. The MFMC algorithm chooses $p \times T$ different one-step transitions to build the MFMC estimator by minimizing the distance $\Delta((y^{l^i_{t-1}}, h(t,y^{l^i_{t-1}})),(x^{l^i_t},u^{l^i_t}))$, so by definition of the k-sparsity of the sample $\mathcal{P}_n$ with $k = pT$, one has $\delta_t^i = \Delta((y^{l^i_{t-1}}, h(t,y^{l^i_{t-1}})),(x^{l^i_t},u^{l^i_t})) \leq \Delta_{pT}^{\mathcal{P}_n}(y^{l^i_{t-1}}, h(t,y^{l^i_{t-1}})) \leq \alpha_{pT}(\mathcal{P}_n)$, which ends the proof.

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

Proof of Theorem 4.1 By definition of $a^h(\tau^i,x_0)$ and $b^h(\tau^i,x_0)$, we have $\forall i \in [\![1,p]\!], \frac{b^h(\tau^i,x_0)+a^h(\tau^i,x_0)}{2} = \sum_{t=0}^{T-1} r^{l^i_t}$. Then, according to Lemmas A.2 and A.3, we have $\forall i \in [\![1,p]\!]$,
$\Big|\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\big[\mathbb{E}[R^h(x_0)|\Omega^{\tau^i}] - \sum_{t=0}^{T-1} r^{l^i_t}\big]\Big| \leq \underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\big[\big|\mathbb{E}[R^h(x_0)|\Omega^{\tau^i}] - \sum_{t=0}^{T-1} r^{l^i_t}\big|\big] \leq C\alpha_{pT}(\mathcal{P}_n)$.
Thus,
$\Big|\frac{1}{p}\sum_{i=1}^p \underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\big[\mathbb{E}[R^h(x_0)|\Omega^{\tau^i}] - \sum_{t=0}^{T-1} r^{l^i_t}\big]\Big| \leq \frac{1}{p}\sum_{i=1}^p \Big|\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\big[\mathbb{E}[R^h(x_0)|\Omega^{\tau^i}] - \sum_{t=0}^{T-1} r^{l^i_t}\big]\Big| \leq C\alpha_{pT}(\mathcal{P}_n)$,
which can be reformulated
$\Big|\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\big[\frac{1}{p}\sum_{i=1}^p \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]\big] - E^h_{p,\mathcal{P}_n}(x_0)\Big| \leq C\alpha_{pT}(\mathcal{P}_n)$, since $\frac{1}{p}\sum_{i=1}^p\sum_{t=0}^{T-1} r^{l^i_t} = \mathfrak{M}_p^h(\tilde{\mathcal{F}}_n,x_0)$.
Since the MFMC algorithm chooses $p \times T$ different one-step transitions, all the $\{w^{l^i_t}\}_{i=1,t=0}^{i=p,t=T-1}$ are i.i.d. according to $p_{\mathcal{W}}(.)$. For all $i \in [\![1,p]\!]$, the law of total expectation gives $\underset{w^{l^i_0},\ldots,w^{l^i_{T-1}}\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\big[\underset{w^{l^i_0},\ldots,w^{l^i_{T-1}}\sim p_{\mathcal{W}}(.)}{\mathbb{E}}[R^h(x_0)|\Omega^{\tau^i}]\big] = \underset{w_0,\ldots,w_{T-1}\sim p_{\mathcal{W}}(.)}{\mathbb{E}}[R^h(x_0)] = J^h(x_0)$. This ends the proof.

B Proof of Theorem 4.2

We first have the following lemma.

Lemma B.1 (Variance of a sum of random variables) *Let* $X_0,\ldots,X_{T-1}$ *be* T *random variables with variances* $\sigma_0^2,\ldots,\sigma_{T-1}^2$ *respectively. Then,*
$Var\big[\sum_{t=0}^{T-1} X_t\big] \leq \big(\sum_{t=0}^{T-1}\sigma_t\big)^2$.

Proof of Lemma B.1 The proof is obtained by induction on the number of random variables using the formula $Cov(X_i,X_j) \leq \sigma_i\sigma_j$, $\forall i,j \in [\![0,T-1]\!]$ which is a straightforward consequence of the Cauchy-Schwarz inequality.

Proof of Theorem 4.2 We denote by $\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n,x_0)$ the random variable $\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n,x_0) = \mathfrak{M}_p^h(\tilde{\mathcal{F}}_n,x_0) - \frac{1}{p}\sum_{i=1}^p \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]$. According to Lemma B.1, we can write

$$V^h_{p,\mathcal{P}_n}(x_0) \leq \Bigg(\sqrt{\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{Var}\big[\frac{1}{p}\sum_{i=1}^p \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]\big]} + \sqrt{\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{Var}\big[\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n,x_0)\big]}\Bigg)^2 \quad (3)$$

Since all the $\{w^{l^i_t}\}_{i=1,t=0}^{i=p,t=T-1}$ are i.i.d. according to $p_{\mathcal{W}}(.)$ (cf proof of Theorem 4.1), the law of total expectation gives

$$\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{Var}\big[\frac{1}{p}\sum_{i=1}^p \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]\big] = \frac{\sigma^2_{R^h}(x_0)}{p} . \quad (4)$$

Now, let us focus on $\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{Var}\big[\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n,x_0)\big]$. By definition, we have $\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n,x_0) = \frac{1}{p}\sum_{i=1}^p\big[\sum_{t=0}^{T-1} r^{l^i_t} - \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]\big]$. Then, according to Lemma B.1, we have

$$\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{Var}\big[\mathfrak{N}_p^h(\tilde{\mathcal{F}}_n,x_0)\big] \leq \frac{1}{p^2}\Bigg(\sum_{i=1}^p \sqrt{\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{Var}\big[\sum_{t=0}^{T-1} r^{l^i_t} - \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]\big]}\Bigg)^2 \quad (5)$$

Then, we can write

$$\begin{aligned}
&\underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{Var}\big[\sum_{t=0}^{T-1} r^{l^i_t} - \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]\big] \\
&\leq \underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\big[(\sum_{t=0}^{T-1} r^{l^i_t} - \mathbb{E}[R^h(x_0)|\Omega^{\tau^i}])^2\big] \\
&\leq \underset{w^1,\ldots,w^n\sim p_{\mathcal{W}}(.)}{\mathbb{E}}\big[\big(a^h(\tau^i,x_0) - b^h(\tau^i,x_0)\big)^2\big] \\
&= \big(a^h(\tau^i,x_0) - b^h(\tau^i,x_0)\big)^2 \\
&\leq 4C^2(\alpha_{pT}(\mathcal{P}_n))^2 . \qquad (6)
\end{aligned}$$

since $\sum_{t=0}^{T-1} r^{l^i_t}$ and $\mathbb{E}[R^h(x_0)|\Omega^{\tau^i}]$ both belong to the interval $[b^h(\tau^i,x_0), a^h(\tau^i,x_0)]$ whose width is bounded by $2C\alpha_{pT}(\mathcal{P}_n)$ according to Lemma A.3.

Using Equations (3), (4), (5) and (6), we have

$$V^h_{p,\mathcal{P}_n}(x_0) \leq \Big(\frac{\sigma_{R^h}(x_0)}{\sqrt{p}} + 2C\alpha_{pT}(\mathcal{P}_n)\Big)^2$$

which ends the proof.