A Tale of Sampling and Estimation in Discounted Reinforcement Learning

Alberto Maria Metelli
Politecnico di Milano

Mirco Mutti
Politecnico di Milano

Marcello Restelli
Politecnico di Milano

Abstract

The most relevant problems in discounted reinforcement learning involve estimating the mean of a function under the stationary distribution of a Markov reward process, such as the expected return in *policy evaluation*, or the policy gradient in *policy optimization*. In practice, these estimates are produced through a finite-horizon episodic sampling, which neglects the mixing properties of the Markov process. It is mostly unclear how this mismatch between the practical and the ideal setting affects the estimation, and the literature lacks a formal study on the pitfalls of episodic sampling, and how to do it optimally. In this paper, we present a minimax lower bound on the discounted mean estimation problem that explicitly connects the estimation error with the mixing properties of the Markov process and the discount factor. Then, we provide a statistical analysis on a set of notable estimators and the corresponding sampling procedures, which includes the finite-horizon estimators often used in practice. Crucially, we show that estimating the mean by directly sampling from the discounted kernel of the Markov process brings compelling statistical properties w.r.t. the alternative estimators, as it matches the lower bound without requiring a careful tuning of the episode horizon.

1 INTRODUCTION

The discounted formulation of the Markov Decision Process (MDP, Puterman, 2014), initially studied in (Blackwell, 1962; Bellman, 1966), established itself as one of the most popular models for Reinforcement Learning (RL, Sutton and Barto, 2018) due to its favorable theoretical tractability and its link with temporal difference learning (Sutton, 1988), a key ingredient behind several successful algorithms (e.g., Watkins and Dayan, 1992; Mnih et al., 2015; Lillicrap et al., 2016; Silver et al., 2016). On a technical level, discounted RL problems are based on the estimation of "exponentially discounted" quantities over an infinite horizon. Specifically, *policy evaluation* requires estimating the γ-discounted value function V_γ^μ of a policy μ:

$$V_\gamma^\mu(s) = \mathbb{E}_\mu \left[\sum_{t=0}^{+\infty} \gamma^t R(s_t, a_t) \,\middle|\, s_0 = s \right], \qquad (1)$$

whereas *policy optimization* involves the estimation of the policy gradient (Sutton et al., 1999):

$$\nabla_\mu V_\gamma^\mu = \mathbb{E}_\mu \left[\sum_{t=0}^{+\infty} \gamma^t \nabla_\mu \log \mu(a_t|s_t) Q_\gamma^\mu(s_t, a_t) \right]. \qquad (2)$$

One can equivalently write those quantities as expectations over the state-action space $\mathbb{E}_{(s,a)\sim\pi_\gamma^\mu}[f(s,a)]$, where $\pi_\gamma^\mu(s,a) := (1-\gamma)\mathbb{E}_\mu[\sum_{t=0}^{+\infty} \gamma^t \mathbf{1}\{s_t = s, a_t = a\}]$ is the γ-*discounted state-action distribution* induced by policy μ. The latter can be seen as the stationary distribution of a suitably defined Markov Chain (MC, Levin and Peres, 2017) obtained from the original MDP by fixing the policy μ and considering, at any step, a reset probability $1-\gamma$ of returning to the initial state. This means that discounted RL is rooted in the mean estimation of a function in an MC. Nonetheless, the latter technical problem has received little attention in the discounted RL literature, which has mostly focused on the pitfalls of common practices (e.g., Thomas, 2014; Lehnert et al., 2018; Nota and Thomas, 2020; Tang et al., 2021; Zhang et al., 2022) and their impact on the learning problem (Jiang et al., 2016; Van Seijen et al., 2019; Amit et al., 2020; Guo et al., 2022). Instead, how to optimally *collect samples* from the MC and how to appropriately perform the *mean estimation* remain mostly obscure.

This paper formally studies the γ-discounted mean estimation in MCs. First, we provide a general formulation of the problem by defining an estimation algorithm as a pairing of a *reset policy*, which is used to make decisions on whether to reset the chain (i.e., re-start from the initial state) at a given step, and an actual *estimator*, from which the estimate is computed on the collected samples. For this notion of estimation algorithm, we introduce a PAC requirement that guarantees a small estimation error with high probability when enough samples are collected. Most importantly,

we derive a *lower bound* on the number of samples required by any estimation algorithm to meet the proposed PAC requirement, which relates the sample complexity to the discount factor γ and the mixing properties of the chain.

Having established the statistical barriers of the problem, we shift our focus toward the properties of practical estimation algorithms. The most common practice in discounted RL is to compute the quantities in Equations (1, 2) through a *finite-horizon* algorithm, i.e., that resets the chain every T steps. This approach is known to suffer from a meaningful *bias* (Thomas, 2014), which can be only partially mitigated with a careful choice of T and correcting factors (though they are seldom used in practice). Alternatively, one can design an *unbiased* estimation algorithm that, at every step, rejects the collected sample with probability γ, otherwise it accepts the sample and resets the chain. However, this *one-sample* estimator has been mostly used as a theoretical tool (e.g., Thomas, 2014; Metelli et al., 2021), since it wastes a large portion of the samples and, thus, suffers from a large variance. Another option is to reset the chain like the *one-sample* estimator, but to compute the estimate over all the collected samples, like the *finite-horizon* estimators. The latter *all-samples* approach introduces some bias by bringing dependent samples, but it mitigates the variance through a greater effective sample size. For the mentioned estimation algorithms, we study their computational properties, derive concentration inequalities, and certify the *all-samples* approach, which practitioners almost neglect, is the one with the best statistical profile, as it results nearly minimax optimal.

Original Contributions In summary, we contribute:

- A formal definition of the problem of γ-discounted mean estimation in Markov chains and its corresponding PAC requirement (Section 3);
- The first minimax lower bound of order $\widetilde{\Omega}(1/\sqrt{N(1-\beta\gamma)})$ on the error of γ-discounted mean estimation in MCs, where N is the number of collected samples and $1 - \beta$ is the *absolute spectral gap* (Levin and Peres, 2017) of the chain (Section 4);
- The analysis of the statistical properties of a family of estimation algorithms that includes the *finite-horizon*, *one-sample*, *all-samples* types (Section 5), in which the *all-samples* approach results in the best statistical profile;
- An empirical evaluation of the mentioned estimation algorithms over simple yet illustrative problems, which uphold the compelling statistical properties of the *all-samples* estimator (Section 6).

Finally, this paper aims to shed light on the statistical barriers of γ-discounted mean estimation in MCs, which stands as the technical bedrock of the discounted RL formulation. As a by-product of this theoretical analysis, the *all-samples* estimation approach emerges as an interesting opportunity for the development of novel practical algorithms for discounted RL supported by compelling statistical properties.

2 PRELIMINARIES

In this section, we introduce the necessary background that will be employed in the subsequent sections of the paper.

Notation Let $\mathcal{X}$ be a set and $\mathfrak{F}$ a σ-algebra on $\mathcal{X}$. We denote with $\mathscr{P}(\mathcal{X})$ the set of probability measures over $(\mathcal{X}, \mathfrak{F})$. We denote with $\mathscr{B}(\mathcal{X})$ the set of $\mathfrak{F}$-measurable real-valued functions. Let $\nu \in \mathscr{P}(\mathcal{X})$ and $f \in \mathscr{B}(\mathcal{X})$, with little abuse of notation, we denote with $\nu : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$ the expectation operator $\nu f = \int_{\mathcal{X}} f(x)\nu(\mathrm{d}x)$. For $p \in [1, \infty)$, we define the $L_p(\pi)$-norm as $\|f\|_{\pi,p}^p = \int_{\mathcal{X}} |f(x)|^p \pi(\mathrm{d}x)$. Let $T : \mathscr{B}(\mathcal{X}) \to \mathscr{B}(\mathcal{X})$ be a linear operator, we define the operator norm as $\|T\|_{\pi,p\to q} = \sup_{\|f\|_{\pi,p}\leqslant 1} \|Tf\|_{\pi,q}$, for $p, q \in [1, \infty)$. Let $\nu, \mu \in \mathscr{P}(\mathcal{X})$, the *chi-square divergence* is defined as $\chi_2(\nu\|\mu) = \left\|(\mathrm{d}\nu/\mathrm{d}\mu - 1)^2\right\|_{\mu}^2$. For $a, b \in \mathbb{N}$ with $a \leqslant b$ we employ the notation $[\![a, b]\!] = \{a, \ldots, b\}$.

Markov Chains A Markov kernel is an $\mathfrak{F}$-measurable function $P : \mathcal{X} \to \mathscr{P}(\mathcal{X})$ mapping every state $x \in \mathcal{X}$ to a probability measure $P(\cdot|x) \in \mathscr{P}(\mathcal{X})$. We denote with $\mathscr{P}(\mathcal{X}, \mathcal{X})$ the set of Markov kernels over $(\mathcal{X}, \mathfrak{F})$. With little abuse of notation, we denote with the same symbol the operator $P : \mathscr{B}(\mathcal{X}) \to \mathscr{B}(\mathcal{X})$ defined as $(Pf)(x) = \int_{\mathcal{X}} f(y)P(\mathrm{d}y|x)$ for $x \in \mathcal{X}$. A probability measure $\pi \in \mathscr{P}(\mathcal{X})$ is *invariant* w.r.t. P if $\pi = \pi P$. Let $\Pi = \mathbf{1}\pi$, we define the *absolute L_2-spectral gap* (Levin and Peres, 2017) as $1 - \beta$, where $\beta = \|P - \Pi\|_{\pi,2\to2}$.

Discounted Sampling Let $\gamma \in [0, 1]$ be a discount factor, $P \in \mathscr{P}(\mathcal{X}, \mathcal{X})$ be a Markov kernel, and $\nu \in \mathscr{P}(\mathcal{X})$ be an initial-state distribution, the γ-*discounted stationary distribution* $\pi_\gamma \in \mathscr{P}(\mathcal{X})$ is defined in several equivalent forms:

$$\pi_\gamma = (1 - \gamma) \sum_{t\in\mathbb{N}} \gamma^t \nu P^t$$

$$= (1 - \gamma)\nu (I - \gamma P)^{-1} = (1 - \gamma)\nu + \gamma\pi_\gamma P. \quad (3)$$

π_γ represents normalized expected count of the times each state is visited, where a visit at time $t \in \mathbb{N}$ counts γ^t. If $\gamma < 1$, π_γ is guaranteed to exist. When $\gamma = 1$, we denote with $\pi = \lim_{\gamma\to1} \pi_\gamma$ the *stationary distribution* of P, if it exists. It is well-known that π_γ is also the stationary distribution of the MC with kernel $P_\gamma = (1 - \gamma)\mathbf{1}\nu + \gamma P$.

3 γ-DISCOUNTED MEAN ESTIMATION

In this section, we formally define the problem of γ-discounted mean estimation in Markov chains. Then, we introduce a general framework for characterizing a broad class of estimators (Section 3.1), and we formally define the PAC requirement to asses their quality (Section 3.2).

Let $\nu \in \mathscr{P}(\mathcal{X})$ be an initial-state distribution and $P \in \mathscr{P}(\mathcal{X}, \mathcal{X})$ be a Markov kernel. Given a discount factor $\gamma \in [0, 1]$ and a measurable function $f \in \mathscr{B}(\mathcal{X})$, our goal
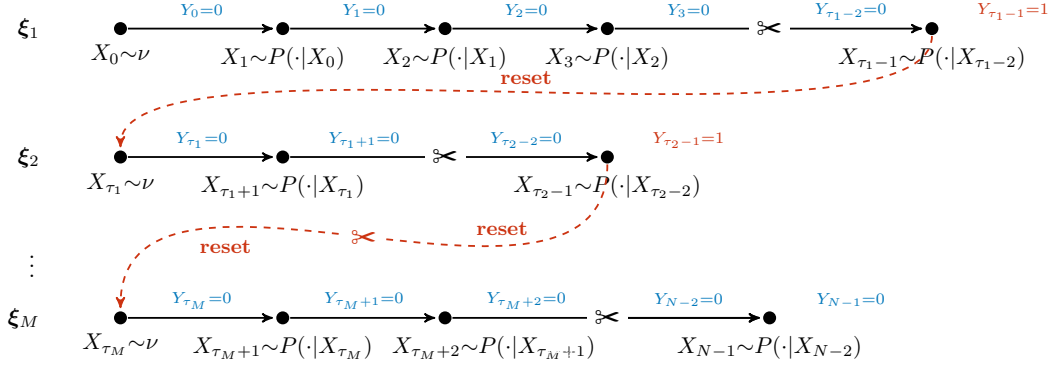
Figure 1: Graphical representation of the sampling process in a Markov chain with a reset policy.

Algorithm 1 Markov chain sampling with reset policy.

Input: Markov kernel P, initial-state distribution ν, discount factor γ, reset policy $\boldsymbol{\rho} = (\rho_t)_{t\in\mathbb{N}}$, number of samples N
Output: dataset H_N

  $H_0 = ()$, $X_0 \sim \nu$
  for $t \in [\![0, N-1]\!]$ **do**
    $Y_t \sim \rho_t(\cdot|\mathcal{H}_t, X_t)$
    $H_{t+1} = H_t \oplus ((X_t, Y_t))$ $\quad$ $\oplus$ denotes concatenation
    if $Y_t = 0$ **then**
      $X_{t+1} \sim P(\cdot|X_t)$
    else
      $X_{t+1} \sim \nu$
    end if
  end for
  return H_N

---

consists in estimating:

$$\pi_\gamma f := \mathbb{E}_{X \sim \pi_\gamma}[f(X)] = \int_{\mathcal{X}} f(x)\pi_\gamma(\mathrm{d}x), \qquad (4)$$

i.e., the expectation of f under the γ-discounted stationary distribution π_γ induced by ν and P, as defined in Equation (3). Furthermore, we introduce the quantity:[1]

$$\sigma_\gamma^2 f := \mathbb{V}\mathrm{ar}_{X \sim \pi_\gamma}[f(X)] = \int_{\mathcal{X}} (f(x) - \pi_\gamma f)^2 \pi_\gamma(\mathrm{d}x),$$

i.e., the variance of f under distribution π_γ. When $\gamma = 1$, we denote with πf and $\sigma^2 f$ the expectation and variance of f under the stationary distribution, when they exist.

3.1 Reset-based Estimation Algorithms

In order to perform a reliable estimation of $\pi_\gamma f$, it is advisable to have the possibility of "resetting" the chain, i.e., to interrupt the natural evolution of the chain based on the Markov kernel P, and restart the simulation from a state sampled from the initial-state distribution ν. For these reasons, we introduce the notion of *reset policy*, i.e., a device that decides whether to reset the chain, based on time, the current history of observed states and reset decisions.

Definition 3.1 (Reset Policy). *A reset policy is a sequence $\boldsymbol{\rho} = (\rho_t)_{t\in\mathbb{N}}$ of functions $\rho_t : \mathcal{H}_t \times \mathcal{X} \to \mathscr{P}(\{0,1\})$ map-*

---

[1]$\pi_\gamma : \mathscr{B}(\mathcal{X}) \to \mathbb{R}$ and $\sigma_\gamma^2 : \mathscr{B}(\mathcal{X}) \to \mathbb{R}_{\geqslant 0}$ act as operators.

ping for every $t \in \mathbb{N}$ a history of past states and resets $H_t = (X_0, Y_0, \ldots, X_{t-1}, Y_{t-1}) \in \mathcal{H}_t = (\mathcal{X} \times \{0,1\})^t$ and the current state $X_t \in \mathcal{X}$, to a probability measure $\rho_t(\cdot|H_t, X_t) \in \mathscr{P}(\{0,1\})$.

Thus, at every time instant $t \in \mathbb{N}$, based on $H_t \in \mathcal{H}_t$ and $X_t \in \mathcal{X}$, we sample the reset decision $Y_t \in \{0,1\}$ from the current reset policy $\rho_t(\cdot|H_t, X_t)$. If $Y_t = 1$, then we reset, i.e., the next state X_{t+1} is sampled from the initial state distribution ν, whereas if $Y_t = 0$ the chain evolution proceeds and X_{t+1} is sampled from the Markov kernel $P(\cdot|X_t)$. The resulting sampling algorithm is reported in Algorithm 1.

Resettable and Resetted Processes Given a reset policy $\boldsymbol{\rho} = (\rho_t)_{t\in\mathbb{N}}$, we can represent the distribution of the next state with the *resettable process*, defined through the resettable Markov kernel $P_\nu : \mathcal{X} \times \{0,1\} \to \mathscr{P}(\mathcal{X})$, defined for every $(X, Y) \in \mathcal{X} \times \{0,1\}$ and measurable set $\mathcal{B} \in \mathfrak{F}$ as:

$$P_\nu(\mathcal{B}|X, Y) = Y \cdot \nu(\mathcal{B}) + (1-Y) \cdot P(\mathcal{B}|X). \quad (5)$$

Suppose we run the process for $N \in \mathbb{N}$ steps, the product measure generating the history H_N is given by $P_{\nu,\boldsymbol{\rho}}^N = \nu \otimes \rho_0 \otimes (\bigotimes_{t=1}^{N-1} P_\nu \otimes \rho_t)$. The sequence of states resulting from applying a reset policy $\boldsymbol{\rho}$ is a non-stationary non-Markovian process, called *resetted process*, whose kernel $P_{\nu,\rho_t,t} : \mathcal{H}_t \times \mathcal{X} \to \mathscr{P}(\mathcal{X})$ is defined for $t \in \mathbb{N}$, history $H \in \mathcal{H}_t$, state $X \in \mathcal{X}$, and measurable set $\mathcal{B} \in \mathfrak{F}$ as:[2]

$$\begin{aligned} P_{\nu,\rho_t,t}(\mathcal{B}|X, H) = &\rho_t(\{1\}|H, X) \cdot \nu(\mathcal{B}) \\ &+ \rho_t(\{0\}|H, X) \cdot P(\mathcal{B}|X). \end{aligned} \quad (6)$$

Trajectories and Horizons We define a *trajectory* as the sequence of states observed between two consecutive resets. The number of trajectories M can be computed in terms of the resets, i.e., $M = 1 + \sum_{t=0}^{N-1} Y_i$. We introduce the time instants τ_i in which a reset is performed as:

$$\tau_i = \begin{cases} 0 & \text{if } i=1 \\ 1+\min\{t\in\mathbb{N} : t \geqslant \tau_{i-1} \wedge Y_t=1\} & \text{if } i\in[\![2,M]\!] \\ N & \text{if } i=M+1 \end{cases} \quad (7)$$

---

[2]If ρ_t is stationary and/or Markovian then $P_{\nu,\rho_t,t}$ is stationary and/or Markovian.

Therefore, for every $i \in [\![M]\!]$, a trajectory is given by $\boldsymbol{\xi}_i = (X_{\tau_i}, \ldots, X_{\tau_{i+1}-1})$, whose horizon is computed as $T_i = \tau_{i+1} - \tau_i$. A graphical representation of the resulting sampling process is provided in Figure 1.

3.2 Estimators and PAC Requirement

In addition to the reset policy $\boldsymbol{\rho}$, to actually define an estimation algorithm, we need an *estimator*, i.e., a function $\widehat{\eta} : \mathcal{H}_N \times \mathscr{B}(\mathcal{X}) \to \mathbb{R}$ that maps a history of observations $H_N \in \mathcal{H}_N$ and a measurable function $f \in \mathscr{B}(\mathcal{X})$ to a real number $\widehat{\eta}(H_N, f) \in \mathbb{R}$. Thus, an *estimation algorithm* is a pair $\mathfrak{A} = (\boldsymbol{\rho}, \widehat{\eta})$. We now introduce the PAC requirement to assess the quality of an estimation algorithm $\mathfrak{A}$.

Definition 3.2 ((ϵ, δ, N)-PAC). *Let $\gamma \in [0,1]$ be a discount factor, let $P \in \mathscr{P}(\mathcal{X}, \mathcal{X})$ be a Markov kernel, let $\nu \in \mathscr{P}(\mathcal{X})$ be an initial-state distribution, and let $f \in \mathscr{B}(\mathcal{X})$ be a measurable function. An estimation algorithm $\mathfrak{A} = (\boldsymbol{\rho}, \widehat{\eta})$ for the γ-discounted mean $\pi_\gamma f$, is (ϵ, δ, N)-PAC if with probability at least $1 - \delta$ it holds that:*

$$|\widehat{\eta}(H_N, f) - \pi_\gamma f| < \epsilon,$$

where $H_N \sim P_{\nu, \boldsymbol{\rho}}^N$ is collected with the reset policy $\boldsymbol{\rho}$.

In the next sections, we first dive into the study of the intrinsic complexity of estimating $\pi_\gamma f$ (Section 4) and, then, we present a handful of practical estimators along with their computational and statistical properties (Section 5).

# 4 MINIMAX LOWER BOUND FOR γ–DISCOUNTED MEAN ESTIMATION

In this section, we prove the first minimax lower bound for the problem of γ–discounted mean estimation in MCs. We first state a lower bound for a general estimation algorithm $\mathfrak{A} = (\boldsymbol{\rho}, \widehat{\eta})$. Then, we report a brief sketch of the proof, which includes how to construct the hard instance, while a complete derivation can be found in Appendix A.1.

Theorem 4.1 (Minimax Lower Bound). *For every discount factor $\gamma \in [0,1]$, sufficiently small confidence δ,[3] number of interactions $N \in \mathbb{N}$, and (ϵ, δ, N)-PAC estimation algorithm $\mathfrak{A} = (\boldsymbol{\rho}, \widehat{\eta})$, there exists a class of Markov kernels $P \in \mathscr{P}(\mathcal{X}, \mathcal{X})$ with absolute spectral gap $1 - \beta \in (0, 1]$, initial-state distributions $\nu \in \mathscr{P}(\mathcal{X})$, measurable function $f \in \mathscr{B}(\mathcal{X})$ such that with probability at least δ it holds that:*

$$|\widehat{\eta}(H_N, f) - \pi_\gamma f| \geqslant \sqrt{\frac{\sigma_\gamma^2 f \cdot \log \frac{1}{2\delta}}{N(1 - \beta\gamma)}},$$

where $H_N \sim P_{\nu, \boldsymbol{\rho}}^N$ is collected with the reset policy $\boldsymbol{\rho}$.

Proof Sketch. The proof is based on the MC construction:

---
[3]The explicit regime for δ is reported in the proof sketch.
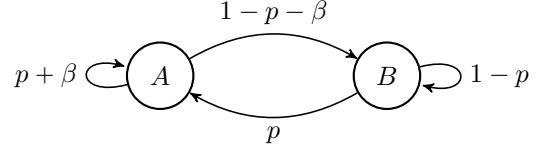
having two states $\mathcal{X} = \{A, B\}$, kernel P parametrized via $\beta \in [0, 1)$ and $p \in (0, 1 - \beta)$, initial state distribution $\nu = (q, 1 - q)$ parametrized via $q \in (0, 1)$. By computing the invariant measure π for the kernel P, it is easy to verify that the MC has spectral gap $1 - \beta$ for every value of p.

We consider a pair of functions f_1, f_{-1} defined as $f_1(B) = f_{-1}(B) = 0$, $f_1(A) = 1$, $f_{-1}(A) = -1$. Crucially, any estimator cannot distinguish the two functions if the state A is never visited. With this intuition, we can lower bound the probability of making an error $\epsilon \in [0, 1]$ through the probability of visiting A. For p and q such that $\pi_\gamma f_1 = \epsilon$ (consequently $\pi_\gamma f_{-1} = -\epsilon$), we can derive:

$$\sup_{\substack{P, \nu, f \\ \text{with spectral gap } 1 - \beta}} \mathbb{P}_{H_N \sim P_{\nu, \boldsymbol{\rho}}^N} (|\widehat{\eta}(H_N, f) - \pi_\gamma f| \geqslant \epsilon)$$

$$\geqslant \frac{1}{2} \mathbb{P}_{H_N \sim P_{\nu, \boldsymbol{\rho}}^N} (\widehat{\eta}(H_N, f_{-1}) = \widehat{\eta}(H_N, f_1))$$

$$\geqslant \frac{1}{2} \min\{1 - q, 1 - p\}^N.$$

Then, we optimize the values of p and q to make the bound tight, and with some algebraic manipulations we get:

$$\frac{1}{2} \min\{1 - q, 1 - p\}^N \geqslant \exp\left(-\frac{\epsilon^2 N(1 - \beta\gamma)}{\sigma_\gamma^2 f}\right),$$

in the regime $\epsilon \in \left[0, \frac{1-\beta}{1-\beta\gamma}\right]$. The statement follows by reformulating the lower bound in terms of deviation ϵ for a small enough $\delta \in \left(0, \frac{1}{2} \exp\left(-\frac{N(1-\beta)^2}{\sigma_\gamma^2 f(1-\beta\gamma)}\right)\right)$. $\square$

The presented minimax lower bound establishes an instance-dependent rate of order $\widetilde{\Omega}(1/\sqrt{N(1 - \beta\gamma)})$ for the deviation in γ–discounted mean estimation, which is the first result that connects the statistical complexity of the problem with both the mixing property of the chain and the discount factor through the term $(1 - \beta\gamma)$. Thus, we can appreciate the role of the spectral gap $1 - \beta$ in governing the complexity of the estimation problem. It is worth noting that, when $\beta = 0$ and the chain mixes instantly making all the collected samples independent, the result reduces to Höeffding's rate (Boucheron et al., 2013) for independent random variables $\widetilde{\Omega}(1/\sqrt{N})$. When $\gamma = 1$, it reduces to Höeffding's rate for general MCs $\widetilde{\Omega}(1/\sqrt{N(1 - \beta)})$ (Fan et al., 2021). Notably, in the latter setting, the problem reduces to the estimation of the mean of a function under the stationary distribution π of an MC. Finally, when $\beta = 1$ and the chain never mixes, the γ–discounted estimation problem is still well-defined for $\gamma < 1$. In Appendix A.1, we report an additional result which shows that resetting the chain is indeed necessary in the latter no-mixing regime.

5 ANALYSIS OF γ–DISCOUNTED MEAN ESTIMATORS

In this section, we analyze four estimation algorithms $\mathfrak{A} = (\boldsymbol{\rho}, \widehat{\eta})$ for the γ–discounted mean $\pi_\gamma f$ from *computational* and *statistical* perspectives. We derive suitable concentration inequalities, compare the estimators and discuss their tightness w.r.t. the provided lower bound. We consider two classes of estimation algorithms, based on the nature of the reset policy: *Fixed-Horizon Reset* (FHR, Section 5.1) and *Adaptive-Horizon Reset* (AHR, Section 5.2). Both classes of estimators assume the knowledge of the discount factor γ but not of the absolute spectral gap $1-\beta$. Table 1 summarizes the properties of the presented estimators. The proofs of the results of this section are reported in Appendix A.2.

5.1 Fixed-Horizon Estimation Algorithms

The *Fixed-Horizon* (FH) estimation algorithms perform a reset action after having experienced a fixed number of transitions $T \in \mathbb{N}$, i.e., generating trajectories with fixed *horizon*. Thus, given N transitions, the number of trajectories is given by $M = \lceil N/T \rceil$, with the last one possibly shorter $T_M = N - (M-1)T$. Thus, the reset policy takes the form $\rho_t^{\text{FHR}}(\cdot|H_t, X_t) = \delta_{\mathbf{1}\{t \bmod T=0\}}$. For the sake of the analysis, we assume that $N \bmod T = 0$, so that all trajectories have the same horizon T.

5.1.1 Computational Properties

The FHR reset policy $\boldsymbol{\rho}^{\text{FHR}}$ is easily parallelizable, as the horizon T is known in advance. Thus, we need $M = N/T$ workers, each collecting a trajectory of T samples, which also corresponds to the time complexity $O(T)$.

5.1.2 Statistical Properties

In the class of FH estimation algorithms, we consider two estimators $\widehat{\eta}$, the *Fixed-Horizon Non-corrected* (FHN) and the *Fixed-Horizon Corrected* (FHC) estimators:

$$\widehat{\eta}_{\text{FHN}}(H_N, f) = \frac{1}{M} \sum_{i=0}^{M-1} (1-\gamma) \sum_{j=0}^{T-1} \gamma^j f(X_{Ti+j}), \quad (8)$$

$$\widehat{\eta}_{\text{FHC}}(H_N, f) = \frac{1}{M} \sum_{i=0}^{M-1} \frac{1-\gamma}{1-\gamma^T} \sum_{j=0}^{T-1} \gamma^j f(X_{Ti+j}). \quad (9)$$

Both estimators are based on a sample average over the $M = N/T$ collected trajectories. The samples of each trajectory are weighted by the discount factor γ raised to a suitable power. The difference between the two estimators lies in the coefficient that multiplies the inner summation. While in the FHN this constant disregards the fact that the summation is limited to the horizon T employing $1-\gamma$ as normalizing constant, the FHC accounts for this by selecting the proper constant $\frac{1-\gamma^T}{1-\gamma} = \sum_{j=0}^{T-1} \gamma^j$. Nonetheless, as we shall see, it is not guaranteed that one estimator

always outperforms the other in all regimes. These estimators are the most widely employed approaches for estimating γ–discounted means in RL (e.g., Deisenroth et al., 2013; Thomas, 2014; Metelli et al., 2020).

Bias Analysis As they truncate each trajectory after T transitions, the FH estimators are affected by a bias, vanishing for large T, which is bounded as follows.

Proposition 5.1 (FH Estimators – Bias). *Let $H_N \sim P_{\nu,\boldsymbol{\rho}}^N$ with the reset policy $\rho_t = \delta_{\mathbf{1}\{t \bmod T=0\}}$, and let $f: \mathcal{X} \to [0,1]$. The bias of the FHN and FHC estimators are upper bounded as:*

$$\underset{H_N \sim P_{\nu,\boldsymbol{\rho}}^N}{\mathbb{B}\text{ias}} [\widehat{\eta}_{FHN}(H_N, f)] \leqslant b_{\gamma,T}^{FHN} := \gamma^T,$$

$$\underset{H_N \sim P_{\nu,\boldsymbol{\rho}}^N}{\mathbb{B}\text{ias}} [\widehat{\eta}_{FHC}(H_N, f)] \leqslant b_{\beta,\gamma,T}^{FHC} := (1-\gamma)\gamma^T \min_{\substack{t_0 \in [\![0,T]\!] \\ s_0 \in \mathbb{N}}} \left\{ \frac{2 - \gamma^{t_0} - \gamma^{s_0}}{1-\gamma} \right.$$

$$\left. + \left(\frac{(\beta\gamma)^{t_0}(1-(\beta\gamma)^{T-t_0})}{(1-\gamma^T)(1-\beta\gamma)} + \frac{(\beta\gamma)^{s_0}\beta^T}{1-\beta\gamma} \right) \sqrt{\chi_2(\nu\|\pi)\sigma^2 f} \right\}.$$

Some observations are in order. First, both estimators are asymptotically unbiased as the horizon $T \to +\infty$. However, none of them is asymptotically unbiased as the budget $N \to +\infty$ (provided that T does not depend on N). Second, the bias of the FHN estimator does not depend on the spectral gap $1-\beta$ of the underlying MC. This is expected, as the normalizing constant generates a scale inhomogeneity, regardless the mixing properties of the MC. Third, the bias of the FHC estimator, instead, depends on the absolute spectral gap $1-\beta$ and on the divergence $\chi_2(\nu\|\pi)$ between the initial-state distribution ν and the stationary distribution π. Thus, in the special case in which $\pi = \nu$ a.s., the bias of the FHC estimator vanishes. Nevertheless, the dependence on β is quite convoluted, and the bound requires an optimization over the auxiliary integer variables t_0 and s_0. Although for the general case the optimization is nontrivial, for the extreme cases $\beta \in \{0, 1\}$, we obtain more interpretable expressions that are reported in the following.

Corollary 5.2 (FHC Estimator – Bias). *Let $H_N \sim P_{\nu,\boldsymbol{\rho}}^N$ with the reset policy $\rho_t = \delta_{\mathbf{1}\{t \bmod T=0\}}$, and let $f: \mathcal{X} \to [0,1]$. The bias of the FHC estimator is upper bounded as:*

• *if $\beta=0$:*

$$b_{0,\gamma,T}^{FHC} = \frac{1-\gamma}{1-\gamma^T}\gamma^T \min\left\{\sqrt{\chi_2(\nu\|\pi)\sigma^2 f}, 1-\gamma^T\right\}; \quad (10)$$

• *if $\beta=1$:*

$$b_{1,\gamma,T}^{FHC} = 2\gamma^T \min\left\{\sqrt{\chi_2(\nu\|\pi)\sigma^2 f}, 1\right\}. \quad (11)$$

Thus, when $\chi_2(\nu\|\pi)\sigma^2 f \gg 1$ the FHC estimator suffers a smaller bias than the FHN one when $\beta = 0$, but, surprisingly larger by a factor 2, when $\beta = 1$. Indeed, when the chain is slowly mixing ($\beta \approx 1$), both will deliver poor estimations and the FHN estimator mitigates this by using a smaller normalization constant. This result, which, to the best of our knowledge, has never appeared in the literature, justifies the use of the non-corrected estimator, especially

Estimator	Computational properties		Statistical properties		
	# parallel workers	Time complexity[*]	Concentration rate [†,‡]		Minimax optimal[§]
			$\beta=0$	$\beta=1$	
FHN	$\dfrac{N}{T}$	T	$\dfrac{1}{\sqrt{N}}$ [¶]	$\dfrac{1}{\sqrt{N(1-\gamma)}}$ [¶]	✗[‖] (✔ for $\beta\in\{0,1\}$)
FHC			$\dfrac{1}{\sqrt{N}}$ [¶]	$\dfrac{1}{\sqrt{N(1-\gamma)}}$ [¶]	✗[‖] (✔ for $\beta\in\{0,1\}$)
OS	M where $M-1\sim\mathrm{Bin}(N-1,1-\gamma)$	$\min\left\{N,\dfrac{\log(N^2/\delta)}{1-\gamma}\right\}$ [‡]	$\dfrac{1}{\sqrt{N(1-\gamma)}}$		✗
AS			$\dfrac{1}{\sqrt{N(1-\beta\gamma)}}$		✔

[*] Big-O.
[†] Big-$\widetilde{O}$.
[‡] With probability at least $1-\delta$.
[§] According to our analysis.
[¶] Selecting $T=O(\log N/\log(1/\gamma))$.
[‖] At least for $\beta\in(\overline{\beta},1)$ with $\overline{\beta}<1$.

Table 1: Summary of the computational and statistical properties of the considered estimators.

when it is known that the underlying MC is slowly mixing.

Concentration Inequalities Let us now move to the derivation of the concentration inequalities for the FH estimators. The technical challenge in this task consists in effectively exploiting the mixing properties of the underlying MC in order to derive tight concentration results. The following provides the general concentration result, which we particularize for specific values of β later.

Theorem 5.3 (FH Estimators – Concentration). *Let $H_N\sim P^N_{\nu,\boldsymbol{\rho}}$ with the reset policy $\rho_t=\delta_{\mathbf{1}\{t\bmod T=0\}}$, and let $f:\mathcal{X}\to[0,1]$. Let us define for $j_0\in[\![0,T]\!]$:*

$$c_{\beta,\gamma}(j_0):=\frac{(\beta\gamma)^{j_0}-(\beta\gamma)^T}{1-\beta\gamma}\sqrt{\chi_2(\nu\|\pi)\sigma^2 f},$$

$$d_{\beta,\gamma}(j_0,\delta):=\sqrt{\frac{8T\left(\log\left(\chi_2(\nu P^{j_0}\|\pi)+1\right)+4\log\frac{2}{\delta}\right)}{N}}$$

$$\times\sqrt{\frac{(1-\gamma^{j_0})^2}{(1-\gamma)^2}+\frac{(1+\beta)(\gamma^{2j_0}-\gamma^{2T})}{(1-\beta)(1-\gamma^2)}}.$$

For every $\delta\in(0,1)$ with probability at least $1-\delta$, it holds:

$$|\widehat{\eta}_{FHN}(H_N,f)-\pi_\gamma f|\leqslant b^{FHN}_{\gamma,T}$$
$$+(1-\gamma)\min_{j_0\in[\![0,T]\!]}\{c_{\beta,\gamma}(j_0)+d_{\beta,\gamma}(j_0,\delta)\}, \quad (12)$$

$$|\widehat{\eta}_{FHC}(H_N,f)-\pi_\gamma f|\leqslant b^{FHC}_{\beta,\gamma,T}$$
$$+\frac{1-\gamma}{1-\gamma^T}\min_{j_0\in[\![0,T]\!]}\{c_{\beta,\gamma}(j_0)+d_{\beta,\gamma}(j_0,\delta)\}. \quad (13)$$

Similarly to the bias case, the resulting expression requires the optimization over a free variable $j_0\in[\![0,T]\!]$. Intuitively, j_0 should be selected (for analysis purpose only) as a function of β. Indeed, for slowly mixing chains ($\beta\approx1$), we should select a small value of j_0 and vice versa. The following corollary provides the order of concentration for the extreme cases $\beta\in\{0,1\}$.

Corollary 5.4 (FH Estimators – Concentration). *Let $H_N\sim P^N_{\nu,\boldsymbol{\rho}}$ with the reset policy $\rho_t=\delta_{\mathbf{1}\{t\bmod T=0\}}$, and let $f:$*

$\mathcal{X}\to[0,1]$. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$, it holds that:[4]

- *if $\beta=0$:*

$$|\widehat{\eta}_{FHN}(H_N,f)-\pi_\gamma f|\leqslant O\left(\gamma^T+\sqrt{\frac{T(1-\gamma)(1-\gamma^T)\log\frac{2}{\delta}}{N}}\right), \quad (14)$$

$$|\widehat{\eta}_{FHC}(H_N,f)-\pi_\gamma f|\leqslant O\left((1-\gamma)\gamma^T+\sqrt{\frac{T(1-\gamma)\log\frac{2}{\delta}}{N(1-\gamma^T)}}\right); \quad (15)$$

- *if $\beta=1$:*

$$|\widehat{\eta}_{FHN}(H_N,f)-\pi_\gamma f|\leqslant O\left(\gamma^T+(1-\gamma^T)\sqrt{\frac{T\log\frac{2}{\delta}}{N}}\right), \quad (16)$$

$$|\widehat{\eta}_{FHC}(H_N,f)-\pi_\gamma f|\leqslant O\left(\gamma^T+\sqrt{\frac{T\log\frac{2}{\delta}}{N}}\right). \quad (17)$$

We note that the FHC estimator outperforms (in the constants, but not in rate) the FHN when $\beta\approx1$, whereas when $\beta\approx0$, the FHN estimator enjoys better concentration.

Remark 5.1 (About Minimax Optimality of the FH Estimators). *A natural question, at this point, is whether the FH estimators match the minimax lower bound of Theorem 4.1. One could, in principle, select a value of the horizon T depending on the spectral gap $1-\beta$ to tighten the confidence bounds. Unfortunately, β is usually unknown in practice. Realistically, one should enforce a value of T that depends on the discount factor γ, and, if necessary, on the confidence δ, and the number of samples N.*

The FH estimators, according to our analysis, do not match the minimax lower bound for general β. When $\beta\in\{0,1\}$, we show in Appendix A.2.3 that the optimal β-independent choice of T is $T^*_\gamma=O((\log N)/(\log(1/\gamma)))$, leading to the rate $\widetilde{O}(1/\sqrt{N})$ for $\beta=0$ and $\widetilde{O}(1/\sqrt{N(1-\gamma)})$ for

---

[4]For interpretability reasons, we ignore the dependence on $\chi_2(\nu\|\pi)\sigma^2 f$.

$\beta = 1$, respectively.[5] *In such regimes, both FH estimators nearly match the minimax lower bound. Nevertheless, in Appendix A.2.4, we show that there exists a regime of large values of β, namely $\beta \in (\overline{\beta}, 1)$ with $\overline{\beta} = (1+\gamma-2\gamma^T)/(1+\gamma-2\gamma^{T+1}) < 1$ for which the concentration rate is at least $\widetilde{\Omega}(1/\sqrt{N(1-\gamma)})$ regardless the value of β (when $0.3 \leqslant \gamma < 1$), not matching the lower bound.*

5.2 Adaptive-Horizon Estimation Algorithms

The *Adaptive-Horizon* (AH) estimation algorithms generate trajectories with possibly different horizons. At $t \in \mathbb{N}$, a Bernoulli random variable with parameter $1-\gamma$ is sampled, leading to the reset policy $\rho_t^{\text{AHR}}(H_t, X_t) = \text{Ber}(1 - \gamma)$. Thus, the horizon T of each trajectory is a random variable too, where $T-1 \sim \text{Geo}(1-\gamma)$ is a geometric distribution.[6]

5.2.1 Computational Properties

In the AHR case, the parallel execution requires computing in advance the horizons $(T_i)_{i \in [\![M]\!]}$ of each trajectory until we ran out of the sample budget N and, subsequently, run in parallel the sample collection of each trajectory. From a technical perspective, characterizing the distribution of the individual T_i is challenging. Indeed, since we need to stop as soon as we reach the budget N, the random variables T_i become dependent. The following result characterizes the distribution of M and the time complexity.

Theorem 5.5 (AH Estimators – Complexity). *Let $H_N \sim P_{\nu,\boldsymbol{\rho}}^N$ with the reset policy $\rho_t^{AHR}(H_t, X_t) = \text{Ber}(1 - \gamma)$. Then, the number of trajectories M is distributed such that $M - 1 \sim \text{Bin}(N - 1, 1 - \gamma)$. Furthermore, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$, the time complexity is bounded as:*

$$\max_{i \in [\![M]\!]} T_i \leqslant O\left(\min\left\{N, \frac{\log\left(N^2/\delta\right)}{1 - \gamma}\right\}\right).$$

Thus, the time complexity is a minimum between N, as no trajectory can be longer than the maximum number of transitions, and a term that grows with γ, since for large γ the trajectories will have, on average, longer lengths.

5.2.2 Statistical Properties

In the family of AH estimators, we analyze the concentration properties of two specific estimation algorithms: *One-Sample* (OS) and *All-Samples* (AS) estimators.

One-Sample Estimator The idea behind the OS estimator is to regard the γ-discounted distribution $\pi_\gamma = \sum_{t \in \mathbb{N}}(1 -$

$\gamma)\gamma^t \nu P^t$ as the mixture of the distributions νP^t with coefficients $(1-\gamma)\gamma^t$. The OS estimator offers a way of generating *independent* samples from π_γ, by retaining the ones right before resetting is performed, i.e., when $Y_t = 1$:

$$\widehat{\eta}_{\text{OS}}(H_N, f) = \frac{1}{M - 1} \sum_{t=0}^{N-1} Y_t f(X_t). \qquad (18)$$

This estimator has been used in (Thomas, 2014; Metelli et al., 2021), mostly for theoretical reasons, being unbiased. The following result provides the concentration.

Theorem 5.6 (OS Estimator – Concentration). *Let $H_N \sim P_{\nu,\boldsymbol{\rho}}^N$ with the reset policy $\rho_t^{AHR}(H_t, X_t) = \text{Ber}(1-\gamma)$, and let $f : \mathcal{X} \to [0, 1]$. For every $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that:*

$$|\widehat{\eta}_{OS}(H_N, f) - \pi_\gamma f| \leqslant \sqrt{\frac{2 \log \frac{8}{\delta}}{N(1 - \gamma)}}.$$

The concentration term is governed by an "effective number of samples" that is $N(1 - \gamma)$. Indeed, the probability of retaining each of the N transitions is $1 - \gamma$. It is worth noting that the concentration bound, as expected, does not depend on the absolute spectral gap $1 - \beta$, since just one sample per trajectory is considered and, consequently, the estimators guarantees vanish as $\gamma \to 1$. Thus, this estimator is not minimax optimal, according to our analysis.

All-Samples Estimator The AS estimator, instead, makes use of all the samples collected from the simulation. Clearly, this choice introduces a new trade-off since we have at our disposal a larger number of samples for estimation, but, unfortunately, within a single trajectory such samples are statistically dependent. Nevertheless, such a dependence is controlled by the mixing properties of the MC. The AS estimator takes the following form:

$$\widehat{\eta}_{\text{AS}}(H_N, f) = \frac{1}{N} \sum_{t=0}^{N-1} f(X_t). \qquad (19)$$

This estimator has been employed in (Konda, 2002; Xu et al., 2020; Eldowa et al., 2022). The following result provides a concentration inequality for the AS estimator that highlights the dependence on the mixing properties.

Theorem 5.7 (AS Estimator – Concentration). *Let $H_N \sim P_{\nu,\boldsymbol{\rho}}^N$ with the reset policy $\rho_t^{AHR}(H_t, X_t) = \text{Ber}(1-\gamma)$, and let $f : \mathcal{X} \to [0, 1]$. For every $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that:*

$$|\widehat{\eta}_{AS}(H_N, f) - \pi_\gamma f| \leqslant \sqrt{\frac{8 \log \frac{2}{\delta} + 4\log(\chi_2(\nu\|\pi_\gamma) + 1)}{N(1 - \beta\gamma)}}.$$

We note the dependence with the spectral gap in $1 - \beta\gamma$. Contrary to the OS estimator, the bound holds even for $\gamma \to 1$. We also observe a logarithmic dependence on the term $\chi_2(\nu\|\pi_\gamma)$ due to the small bias introduced by the sampling procedure. Most importantly, the AS estimator nearly matches the minimax lower bound of Theorem 4.1.

---

[5]Any choice of T independent of N (including the widely employed "effective horizon" $T = 1/(1 - \gamma)$) will never lead to a consistent estimator, since the bias will not vanish as $N \to +\infty$.

[6]In a different perspective, one may first sample $T - 1 \sim \text{Geo}(1 - \gamma)$ and then simulate a trajectory of horizon T.
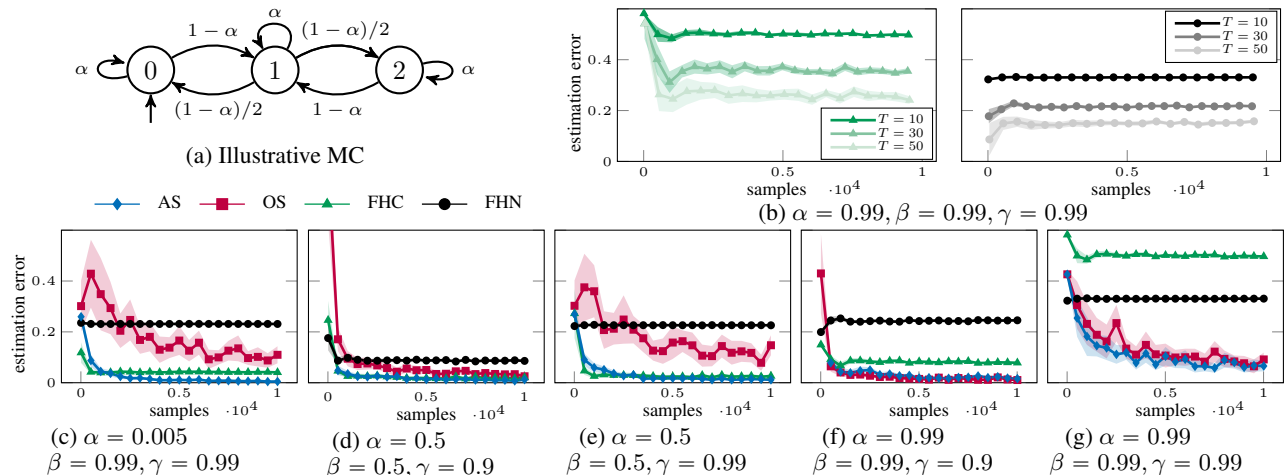
(a) Illustrative MC

(b) $\alpha = 0.99, \beta = 0.99, \gamma = 0.99$

(c) $\alpha = 0.005$
$\beta = 0.99, \gamma = 0.99$

(d) $\alpha = 0.5$
$\beta = 0.5, \gamma = 0.9$

(e) $\alpha = 0.5$
$\beta = 0.5, \gamma = 0.99$

(f) $\alpha = 0.99$
$\beta = 0.99, \gamma = 0.9$

(g) $\alpha = 0.99$
$\beta = 0.99, \gamma = 0.99$

Figure 2: γ–discounted mean estimation of the function $f(x) = (1, -1, 2)$ over the MC depicted in **(a)**. For each combination of parameter α and discount γ, we report the estimation error of the OS, AS, FHC, FHN estimators **(c, d, e, f, g)**. For the FHC, FHN, we provide a finer analysis on the impact of T **(b)**. We report average and 95% c.i. over 20 runs.

6 NUMERICAL VALIDATION

In this section, we confront the γ-discounted mean estimators presented in Section 5 through numerical simulations to both support and complement the analysis of their statistical properties. To this end, we consider an illustrative family of MCs parametrized by α (Figure 2a). The parameter α allows controlling the mixing properties of the chain. If we set α close to either 0 or 1 we get a slow-mixing chain (large β), in which every state is nearly transient or absorbing respectively. α close to $1/2$ gives a fast-mixing chain instead (small β). In this setting, we consider the problem of estimating the mean of the function $f(x) = (1, -1, 2)$ in different discounting regimes, namely $\gamma \in \{0.9, 0.99\}$, where the performance of each estimator $\widehat{\eta}$ is measured in terms of the corresponding estimation error $|\widehat{\eta} - \pi_\gamma f|$.

In Figure 2, we report the results of the numerical analysis. As a further testament of its compelling statistical properties, the AS estimator dominates the alternatives, achieving the smallest estimation error in every mixing-discounting regime. Interestingly, the unbiased OS estimator fails to quickly converge to the true mean with high discounting (Figures 2c 2e 2g), which is likely caused by its inherently large variance. The finite-horizon estimators FHC and FHN show a significant bias instead, despite an overall stable behavior. Although the corrected estimator FHC outperforms (as expected) the non-corrected FHN in most of the regimes (see Figures 2c-2f), the correction can skyrocket the bias in some unfortunate settings (see Figure 2g). This is particularly underwhelming as we cannot trust FHC as a default option even when committed to a finite-horizon estimation algorithm. Finally, in Figure 2b we provide a finer analysis of the finite-horizon estimators for different horizons T. Unsurprisingly, in a slow-mixing regime (note $\beta = 0.99$), increasing T benefits the overall quality of the

estimates for both FHC and FHN, as the bias is visually reduced at the cost of a slightly increased instability.

7 CONCLUSIONS

In this paper, we have studied the problem of estimating the mean of a function under the γ–discounted stationary distribution of an MC. We have formulated this problem with a general and flexible framework and then analyzed its intrinsic complexity through a minimax lower bound. Finally, we have considered two classes of estimation algorithms, for which we provided a study of computation and statistical properties, as well as a numerical validation.

The aim of this paper is far from being a theoretical detour, as we believe that our contribution has significant practical implications in discounted RL. Especially, the all-samples estimator resulted in the best statistical profile among the considered alternatives, while still supporting parallel sampling. This signals an avenue to develop improved "deep reinforcement learning" (François-Lavet et al., 2018) algorithms based on sampling from the discounted kernel. Other interesting future directions include the study of the γ-discounted mean estimation for inhomogeneous functions, which is akin to the practical implementations of Q-learning (Watkins and Dayan, 1992), the estimation of other functionals beyond the expectation (Chandak et al., 2021), and extending our analysis to generalized notions of discount (Yoshida et al., 2013; François-Lavet et al., 2015; Pitis, 2019; Fedus et al., 2019; Tang et al., 2021).

Finally, our results can be of independent interest in the MC literature while bridging fundamental problems in discounted RL and concentration inequalities for MCs (Samson, 2000; Glynn and Ormoneit, 2002; León and Perron, 2004; Kontorovich and Ramanan, 2008; Paulin, 2015).

References

Amit, R., Meir, R., and Ciosek, K. (2020). Discount factor as a regularizer in reinforcement learning. In *International Conference on Machine Learning*.

Bellman, R. (1966). Dynamic programming. *Science*, 153(3731):34–37.

Blackwell, D. (1962). Discrete dynamic programming. *The Annals of Mathematical Statistics*, pages 719–726.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Chandak, Y., Niekum, S., da Silva, B., Learned-Miller, E., Brunskill, E., and Thomas, P. S. (2021). Universal off-policy evaluation. In *Advances in Neural Information Processing Systems*.

Chatzigeorgiou, I. (2013). Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Commun. Lett.*, 17(8):1505–1508.

Deisenroth, M. P., Neumann, G., and Peters, J. (2013). A survey on policy search for robotics. *Found. Trends Robotics*, 2(1-2):1–142.

Eldowa, K., Bisi, L., and Restelli, M. (2022). Finite sample analysis of mean-volatility actor-critic for risk-averse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*.

Fan, J., Jiang, B., and Sun, Q. (2021). Hoeffding's inequality for general markov chains and its applications to statistical learning. *Journal of Machine Learning Research*, 22(139):1–35.

Fedus, W., Gelada, C., Bengio, Y., Bellemare, M. G., and Larochelle, H. (2019). Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*.

François-Lavet, V., Fonteneau, R., and Ernst, D. (2015). How to discount deep reinforcement learning: Towards new dynamic strategies. *arXiv preprint arXiv:1512.02011*.

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., Pineau, J., et al. (2018). An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 11(3-4):219–354.

Glynn, P. W. and Ormoneit, D. (2002). Hoeffding's inequality for uniformly ergodic Markov chains. *Statistics & Probability Letters*, 56(2):143–146.

Guo, X., Hu, A., and Zhang, J. (2022). Theoretical guarantees of fictitious discount algorithms for episodic reinforcement learning and global convergence of policy gradient methods. In *AAAI Conference on Artificial Intelligence*.

Haveliwala, T. and Kamvar, S. (2003). The second eigenvalue of the google matrix. Technical report, Stanford.

Jiang, N., Singh, S., and Tewari, A. (2016). On structural properties of mdps that bound loss due to shallow planning. In *International Joint Conference on Artificial Intelligence*.

Konda, V. (2002). *Actor-critic algorithms*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Kontorovich, L. A. and Ramanan, K. (2008). Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158.

Lehnert, L., Laroche, R., and van Seijen, H. (2018). On value function representation of long horizon problems. In *AAAI Conference on Artificial Intelligence*.

León, C. A. and Perron, F. (2004). Optimal Hoeffding bounds for discrete reversible Markov chains. *The Annals of Applied Probability*, 14(2):958–970.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2016). Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*.

Metelli, A. M., Pirotta, M., Calandriello, D., and Restelli, M. (2021). Safe policy iteration: A monotonically improving approximate policy iteration approach. *Journal of Machine Learning Research*, 22:97–1.

Metelli, A. M., Pirotta, M., and Restelli, M. (2020). On the use of the policy gradient and hessian in inverse reinforcement learning. *Intelligenza Artificiale*, 14(1):117–150.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Nota, C. and Thomas, P. S. (2020). Is the policy gradient a gradient? In *International Conference on Autonomous Agents and MultiAgent Systems*.

Paulin, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32.

Pitis, S. (2019). Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *AAAI Conference on Artificial Intelligence*.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Samson, P.-M. (2000). Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*.

Tang, Y., Rowland, M., Munos, R., and Valko, M. (2021). Taylor expansion of discount factors. In *International Conference on Machine Learning*.

Thomas, P. (2014). Bias in natural actor-critic algorithms. In *International Conference on Machine Learning*.

Van Seijen, H., Fatemi, M., and Tavakoli, A. (2019). Using a logarithmic mapping to enable lower discount factors in reinforcement learning. In *Advances in Neural Information Processing Systems*.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3):279–292.

Wilkinson, J. (1971). The algebraic eigenvalue problem. In *Handbook for Automatic Computation, Volume II, Linear Algebra*. Springer-Verlag New York.

Xu, T., Wang, Z., and Liang, Y. (2020). Improving sample complexity bounds for (natural) actor-critic algorithms. In *Advances in Neural Information Processing Systems*.

Yoshida, N., Uchibe, E., and Doya, K. (2013). Reinforcement learning with state-dependent discount factor. In *Joint International Conference on Development and Learning and Epigenetic Robotics*.

Zhang, S., Laroche, R., van Seijen, H., Whiteson, S., and Tachet des Combes, R. (2022). A deeper look at discounting mismatch in actor-critic algorithms. In *International Conference on Autonomous Agents and Multi-Agent Systems*.
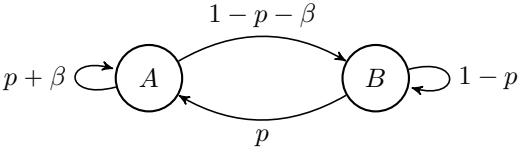
A PROOFS

A.1 Proofs of Section 4

Theorem 4.1 (Minimax Lower Bound). *For every discount factor $\gamma \in [0,1]$, sufficiently small confidence δ, number of interactions $N \in \mathbb{N}$, and (ϵ, δ, N)-PAC estimation algorithm $\mathfrak{A} = (\boldsymbol{\rho}, \widehat{\eta})$, there exists a class of Markov kernels $P \in \mathscr{P}(\mathcal{X}, \mathcal{X})$ with absolute spectral gap $1 - \beta \in (0,1]$, initial-state distributions $\nu \in \mathscr{P}(\mathcal{X})$, measurable function $f \in \mathscr{B}(\mathcal{X})$ such that with probability at least δ it holds that:*

$$|\widehat{\eta}(H_N, f) - \pi_\gamma f| \geqslant \sqrt{\frac{\sigma_\gamma^2 f \cdot \log \frac{1}{2\delta}}{N(1 - \beta\gamma)}},$$

where $H_N \sim P_{\nu,\boldsymbol{\rho}}^N$ is collected with the reset policy $\boldsymbol{\rho}$.

Proof. The proof is articulated in four steps.

First step: Chain construction We consider a 2-states MC with state space $\mathcal{X} = \{A, B\}$ whose kernel is parametrized via p and β:

$$P = \begin{pmatrix} p + \beta & 1 - p - \beta \\ p & 1 - p \end{pmatrix},$$

where $\beta \in [0, 1)$ and $p \in (0, 1 - \beta)$. The kernel P admits eigenvalues $\{1, \beta\}$ and has a unique invariant measure $\pi = \left(\frac{p}{1-\beta}, 1 - \frac{p}{1-\beta}\right)$. We immediately verify that $\Pi^T \odot P$ is symmetric, where $\Pi = \mathbf{1}\pi$ and $\odot$ denotes the Hadamard product. Consequently, the MC is reversible and, thus, its spectral gap is $1 - \beta$. We consider as initial state distribution $\nu = (q, 1 - q)$ parametrized by $q \in (0, 1)$. The specific values of p and q will be specified later in the proof.

Consider two functions $f_1, f_{-1} : \mathcal{X} \in [-1, 1]$ defined as: $f_s(A) = s$ and $f_s(B) = 0$ for $s \in \{-1, 1\}$. Simple calculations allow to show that the corresponding expectations and variances, under the γ-discounted stationary distributions are given for $s \in \{-1, 1\}$ by:

$$\pi_\gamma f_s = \sum_{x \in \mathcal{X}} \pi_\gamma(x) f_s(x) = s \frac{(1 - \gamma)q + \gamma p}{1 - \beta\gamma},$$
$$\sigma_f^2 f_s = \pi_\gamma (f_s - \mathbf{1}\pi_\gamma f)^2 = \pi_\gamma f_s (1 - \pi_\gamma f_s).$$

Second step: Lower bounding the probability of deviation We now proceed at lower bounding the probability of making an error larger than ϵ, with $\epsilon \in [0, 1]$. The intuition is that we need to compute the probability not to distinguish the two instances, i.e., never visiting state A.

For any values of p and q such that $\pi_\gamma f_1 = \epsilon$, considering a generic estimator $(\boldsymbol{\rho}, \widehat{\eta})$, we have:

$$\sup_{\substack{P,\nu,f \\ \text{with spectral gap } \beta}} \mathbb{P}_{H_N \sim P_{\nu,\boldsymbol{\rho}}^N} (|\widehat{\eta}(H_N, f) - \pi_\gamma f| \geqslant \epsilon) \geqslant \max_{s \in \{-1,1\}} \mathbb{P}_{H_N \sim P_{\nu,\boldsymbol{\rho}}^N} (|\widehat{\eta}(H_N, f_s) - \pi_\gamma f_s| \geqslant \epsilon) \tag{20}$$

$$\geqslant \frac{1}{2} \left(\mathbb{P}_{H_N \sim P_{\nu,\boldsymbol{\rho}}^N} (|\widehat{\eta}(H_N, f_{-1}) - \pi_\gamma f_{-1}| \geqslant \epsilon) + \mathbb{P}_{H_N \sim P_{\nu,\boldsymbol{\rho}}^N} (|\widehat{\eta}(H_N f_1) - \pi_\gamma f_1| \geqslant \epsilon) \right) \tag{21}$$

$$\geqslant \frac{1}{2} \mathbb{P}_{H_N \sim P_{\nu,\boldsymbol{\rho}}^N} (|\widehat{\eta}(H_N, f_{-1}) - \pi_\gamma f_{-1}| \geqslant \epsilon \vee |\widehat{\eta}(H_N, f_1) - \pi_\gamma f_1| \geqslant \epsilon) \tag{22}$$

$$\geqslant \frac{1}{2} \mathbb{P}_{H_N \sim P_{\nu,\boldsymbol{\rho}}^N} (\widehat{\eta}(H_N, f_{-1}) = \widehat{\eta}(H_N, f_1)) \tag{23}$$

$$\geqslant \frac{1}{2} \mathbb{P}_{H_N \sim P_{\nu,\boldsymbol{\rho}}^N} (\forall t \in \{0, \ldots, N - 1\} : X_t = B) \tag{24}$$

$$= \frac{1}{2}\nu(\{B\}) \min\{\nu(\{B\}), P(\{B\}|B)\}^{N-1} \geqslant \frac{1}{2}\min\{1-q, 1-p\}^N, \tag{25}$$

where in line (27) we exploited the inequality $\max\{x, y\} \geqslant \frac{1}{2}(x+y)$, in line (28) we employed a union bound, in line (29) we used the fact that $\pi_\gamma f_s = s\epsilon$ by assumption and consequently the event $\{\widehat{\eta}(H_N, f_{-1}) = \widehat{\eta}(H_N, f_1)\}$ is included in the event $\{|\widehat{\eta}(H_N, f_{-1}) - \pi_\gamma f_{-1}| > \epsilon\} \cup \{|\widehat{\eta}(H_N, f_1) - \pi_\gamma f_{-1}| > \epsilon\}$. In line (30) comes from the observation that in order to have equal values of the estimators we must not distinguish the two chain instances, that in turn happens only when we never visit state A. Finally, line (31) follows by taking the minimum probability for never landing to state A depending on both reset and transition probability.

For any values of p and q such that $\pi_\gamma f_1 = \epsilon$, considering a generic estimator $(\boldsymbol{\rho}, \widehat{\eta})$, we have:

$$\sup_{\substack{P,\nu,f \\ \text{with spectral gap } \beta}} \mathbb{P}_{H_N \sim P^N_{\nu,\rho}} \left(|\widehat{\eta}(H_N, f) - \pi_\gamma f| \geqslant \epsilon\right) \geqslant \max_{s \in \{-1,1\}} \mathbb{P}_{H_N \sim P^N_{\nu,\rho}} \left(|\widehat{\eta}(H_N, f_s) - \pi_\gamma f_s| \geqslant \epsilon\right) \tag{26}$$

$$\geqslant \frac{1}{2}\left(\mathbb{P}_{H_N \sim P^N_{\nu,\rho}} \left(|\widehat{\eta}(H_N, f_{-1}) - \pi_\gamma f_{-1}| \geqslant \epsilon\right) + \mathbb{P}_{H_N \sim P^N_{\nu,\rho}} \left(|\widehat{\eta}(H_N f_1) - \pi_\gamma f_1| \geqslant \epsilon\right)\right) \tag{27}$$

$$\geqslant \frac{1}{2}\mathbb{P}_{H_N \sim P^N_{\nu,\rho}} \left(|\widehat{\eta}(H_N, f_{-1}) - \pi_\gamma f_{-1}| \geqslant \epsilon \vee |\widehat{\eta}(H_N, f_1) - \pi_\gamma f_1| \geqslant \epsilon\right) \tag{28}$$

$$\geqslant \frac{1}{2}\mathbb{P}_{H_N \sim P^N_{\nu,\rho}} \left(\widehat{\eta}(H_N, f_{-1}) = \widehat{\eta}(H_N, f_1)\right) \tag{29}$$

$$\geqslant \frac{1}{2}\mathbb{P}_{H_N \sim P^N_{\nu,\rho}} \left(\forall t \in \{0, \dots, N-1\} : X_t = B\right) \tag{30}$$

$$= \frac{1}{2}\nu(\{B\}) \min\{\nu(\{B\}), P(\{B\}|B)\}^{N-1} \geqslant \frac{1}{2}\min\{1-q, 1-p\}^N, \tag{31}$$

where in line (27) we exploited the inequality $\max\{x, y\} \geqslant \frac{1}{2}(x+y)$, in line (28) we employed a union bound, in line (29) we used the fact that $\pi_\gamma f_s = s\epsilon$ by assumption and consequently the event $\{\widehat{\eta}(H_N, f_{-1}) = \widehat{\eta}(H_N, f_1)\}$ is included in the event $\{|\widehat{\eta}(H_N, f_{-1}) - \pi_\gamma f_{-1}| > \epsilon\} \cup \{|\widehat{\eta}(H_N, f_1) - \pi_\gamma f_{-1}| > \epsilon\}$. In line (30) comes from the observation that in order to have equal values of the estimators we must not distinguish the two chain instances, that in turn happens only when we never visit state A. Finally, line (31) follows by taking the minimum probability for never landing to state A depending on both reset and transition probability.

Third step: Tightening the bound Now, we need to compute the values of p and q in order to make the bound as tight as possible while fulfilling all the constraints. This leads to the optimization problem:

$$\max_{p,q} \min\{1-q, 1-p\}$$
$$\text{s.t. } 0 < p < 1 - \beta$$
$$0 < q < 1$$
$$\frac{(1-\gamma)q + \gamma p}{1 - \beta\gamma} = \epsilon.$$

First of all, we exploit the constraint with equality to express q as a function of p, i.e., $q = \frac{\epsilon(1-\beta\gamma)}{1-\gamma} - \frac{\gamma p}{1-\gamma}$. Now, we consider the two cases:

Case 1: $p \leqslant q$ In this case, the $\min$ in the objective function reduces to $1 - q = 1 - \frac{\epsilon(1-\beta\gamma)}{1-\gamma} + \frac{\gamma p}{1-\gamma}$ that is maximized by taking the maximum value of p fulfilling the constraints:

$$\max_p 1 - \frac{\epsilon(1-\beta\gamma)}{1-\gamma} + \frac{\gamma p}{1-\gamma}$$
$$\text{s.t. } 0 < p < 1 - \beta$$
$$0 < q < 1 \implies 0 < \frac{\epsilon(1-\beta\gamma)}{1-\gamma} - \frac{\gamma p}{1-\gamma} < 1$$
$$p \leqslant q \implies p \leqslant \frac{\epsilon(1-\beta\gamma)}{1-\gamma} - \frac{\gamma p}{1-\gamma}$$

This leads to:

$$p = \begin{cases} \epsilon(1 - \beta\gamma) & \text{if } \epsilon \in \left[0, \frac{1-\beta}{1-\beta\gamma}\right] \\ 1 - \beta & \text{if } \epsilon \in \left(\frac{1-\beta}{1-\beta\gamma}, 1\right] \end{cases} \implies 1 - q = \begin{cases} 1 - \epsilon(1 - \beta\gamma) & \text{if } \epsilon \in \left[0, \frac{1-\beta}{1-\beta\gamma}\right] \\ \frac{(1-\beta\gamma)(1-\epsilon)}{1-\gamma} & \text{if } \epsilon \in \left(\frac{1-\beta}{1-\beta\gamma}, 1\right] \end{cases}.$$

Case 2: $p > q$ In this case, the $\min$ in the objective function reduces to $1 - p$ that is maximized by taking the minimum value of p fulfilling the constraints:

$$\max_p 1 - p$$

$$\text{s.t. } 0 < p < 1 - \beta$$

$$0 < q < 1 \implies 0 < \frac{\epsilon(1 - \beta\gamma)}{1 - \gamma} - \frac{\gamma p}{1 - \gamma} < 1$$

$$p > q \implies p > \frac{\epsilon(1 - \beta\gamma)}{1 - \gamma} - \frac{\gamma p}{1 - \gamma}$$

The problem is feasible only when $\epsilon \in \left[0, \frac{1-\beta}{1-\beta\gamma}\right]$. In such a case, we have:

$$p = \epsilon(1 - \beta\gamma) \implies 1 - q = 1 - \epsilon(1 - \beta\gamma) \quad \text{if } \epsilon \in \left[0, \frac{1-\beta}{1-\beta\gamma}\right].$$

Fourth step: Algebraic manipulation We now proceed at performing some manipulation to get more interpretable result. In the small-ϵ regime, we have:

$$(1 - \epsilon(1 - \beta\gamma))^N \geqslant \exp\left(-\frac{\epsilon N(1 - \beta\gamma)}{1 - \epsilon(1 - \beta\gamma)}\right) \tag{32}$$

$$= \exp\left(-\frac{\epsilon^2 N(1 - \beta\gamma)}{\sigma_\gamma^2 f} \cdot \frac{1 - \epsilon}{1 - \epsilon(1 - \beta\gamma)}\right) \tag{33}$$

$$\geqslant \exp\left(-\frac{\epsilon^2 N(1 - \beta\gamma)}{\sigma_\gamma^2 f}\right), \tag{34}$$

where we exploited the inequality $1 - x \geqslant \exp(-x/(1 - x))$ in line (32), and the fact that $\sigma^2 f = \epsilon(1 - \epsilon)$ in line (33). Following similar steps for the large-ϵ regime, we have:

$$\left(\frac{(1 - \beta\gamma)(1 - \epsilon)}{1 - \gamma}\right)^N \geqslant \exp\left(-N \frac{1 - \frac{(1-\beta\gamma)(1-\epsilon)}{1-\gamma}}{\frac{(1-\beta\gamma)(1-\epsilon)}{1-\gamma}}\right)$$

$$= \exp\left(-\frac{N\epsilon\left(1 - \frac{(1-\beta\gamma)(1-\epsilon)}{1-\gamma}\right)}{\sigma_\gamma^2 f} \cdot \frac{1 - \epsilon}{\frac{(1-\beta\gamma)(1-\epsilon)}{1-\gamma}}\right)$$

$$= \exp\left(-\frac{N\epsilon\left(\epsilon(1 - \beta\gamma) - \gamma(1 - \beta)\right)}{\sigma_\gamma^2 f \cdot (1 - \beta\gamma)}\right)$$

$$\geqslant \exp\left(-\frac{N\epsilon^2}{\sigma_\gamma^2 f}\right).$$

Finally, we can reformulate the previous results on the confidence δ in terms of deviation ϵ, such that we have with probability at least $1 - \delta$

$$\inf_{\boldsymbol{\rho}, \widehat{\eta}} \sup_{\substack{P, \nu, f \\ \text{with spectral gap } 1 - \beta}} |\widehat{\eta}(H_N, f) - \pi_\gamma f| \geqslant \begin{cases} \sqrt{\frac{\sigma_\gamma^2 f \log \frac{1}{2\delta}}{N(1 - \beta\gamma)}} & \text{if } \delta \in \left(0, \frac{1}{2} \exp\left(-\frac{N(1-\beta)^2}{\sigma_\gamma^2 f(1-\beta\gamma)}\right)\right) \\ \sqrt{\frac{\sigma_\gamma^2 f \log \frac{1}{2\delta}}{N}} & \text{otherwise} \end{cases},$$

which concludes the proof. For the sake of clarity, we only report the most meaningful high-confidence regime in the theorem statement. $\square$

The following result shows that the reset is unavoidable, at least, for the case in which the underlying MC does not mix, i.e., when $\beta = 1$.

Theorem A.1 (Reset is Unavoidable). *For any non-reset policy, i.e., $\boldsymbol{\rho} = (\rho_t)_{t \in \mathbb{N}}$ such that $\rho(\cdot | H_t, X_t) = \delta_0(\cdot)$ it holds:*

$$\inf_{\widehat{\eta}} \sup_{\substack{P, \nu, f \\ \text{with spectral gap } 1 - \beta}} \mathbb{P}_{H_N \sim P^N_{\nu, \delta_0}} \left(|\widehat{\eta}(H_N, f) - \pi_\gamma f| > \frac{1}{2} \right) \geqslant \frac{1}{4} \left(\frac{1 + \beta}{2} \right)^{N-1}.$$

Proof. We consider a 2-states MC $\mathcal{X} = \{A, B\}$, with kernel:

$$P = \begin{pmatrix} \frac{1+\beta}{2} & \frac{1-\beta}{2} \\ \frac{1-\beta}{2} & \frac{1+\beta}{2} \end{pmatrix}.$$

It is easy to see that the spectral gap is β. Consider two functions $f_1, f_{-1} : \mathcal{X} \in [-1, 1]$ defined as: $f_s(A) = s$ and $f_s(B) = 0$ for $s \in \{-1, 1\}$. Consider the initial state distribution $\nu = (1/2, 1/2)$. It is simple to show that $\pi_\gamma = (1/2, 1/2)$ and, consequently $\pi_\gamma f_s = s/2$ for $s \in \{-1, 1\}$. Consider now a non-reset policy, it holds that:

$$\sup_{\substack{P, \nu, f \\ \text{with spectral gap } 1 - \beta}} \mathbb{P}_{H_N \sim P^N_{\nu, \delta_0}} \left(|\widehat{\eta}(H_N) - \pi_\gamma f| > \frac{1}{2} \right)$$

$$\geqslant \max_{f_s : s \in \{-1, 1\}} \mathbb{P}_{H_N \sim P^N_{\nu, \delta_0}} \left(|\widehat{\eta}(H_N) - \pi_\gamma f_s| > \frac{1}{2} \right)$$

$$\geqslant \frac{1}{2} \left(\mathbb{P}_{H_N \sim P^N_{\nu, \delta_0}} \left(|\widehat{\eta}(H_N) - \pi_\gamma f_{-1}| > \frac{1}{2} \right) + \mathbb{P}_{H_N \sim P^N_{\nu, \delta_0}} \left(|\widehat{\eta}(H_N) - \pi_\gamma f_1| > \frac{1}{2} \right) \right)$$

$$\geqslant \frac{1}{2} \mathbb{P}_{H_N \sim P^N_{\nu, \delta_0}} \left(|\widehat{\eta}(H_N, f_{-1}) - \pi_\gamma f_{-1}| > \frac{1}{2} \vee |\widehat{\eta}(H_N, f_1) - \pi_\gamma f_1| > \frac{1}{2} \right)$$

$$\geqslant \frac{1}{2} \mathbb{P}_{H_N \sim P^N_{\nu, \delta_0}} \left(\widehat{\eta}(H_N, f_{-1}) = \widehat{\eta}(H_N, f_1) \right)$$

$$\geqslant \frac{1}{2} \mathbb{P}_{H_N \sim P^N_{\nu, \delta_0}} \left(\forall t \in \{0, \dots, N-1\} : X_t = B \right) = \frac{1}{4} \left(\frac{1 + \beta}{2} \right)^{N-1}.$$

$\square$

The latter result implies that, when $\beta = 1$ and the chain never mixes, we have

$$\inf_{\boldsymbol{\rho}, \widehat{\eta}} \sup_{\substack{P, \nu, f \\ \text{with spectral gap } 0}} \mathbb{P}_{H_N \sim P^N_{\nu, \delta_0}} \left(|\widehat{\eta}(H_N, f) - \pi_\gamma f| > \frac{1}{2} \right) \geqslant \frac{1}{4},$$

showing that reset is actually necessary.

A.2 Proofs of Section 5

A.2.1 Fixed-Horizon Estimation Algorithms

Bias Analysis

Proposition 5.1 (FH Estimators – Bias). *Let $H_N \sim P^N_{\nu, \boldsymbol{\rho}}$ with the reset policy $\rho_t = \delta_{\mathbf{1}\{t \bmod T = 0\}}$, and let $f : \mathcal{X} \to [0, 1]$. The bias of the FHN and FHC estimators are upper bounded as:*

$$\underset{H_N \sim P^N_{\nu, \rho}}{\mathrm{Bias}} [\widehat{\eta}_{FHN}(H_N, f)] \leqslant b^{FHN}_{\gamma, T} := \gamma^T,$$

$$\underset{H_N \sim P^N_{\nu, \rho}}{\mathrm{Bias}} [\widehat{\eta}_{FHC}(H_N, f)] \leqslant b^{FHC}_{\beta, \gamma, T} := (1 - \gamma)\gamma^T \min_{\substack{t_0 \in [\![0, T]\!] \\ s_0 \in \mathbb{N}}} \left\{ \frac{2 - \gamma^{t_0} - \gamma^{s_0}}{1 - \gamma} \right.$$

$$+ \left. \left(\frac{(\beta\gamma)^{t_0}(1 - (\beta\gamma)^{T - t_0})}{(1 - \gamma^T)(1 - \beta\gamma)} + \frac{(\beta\gamma)^{s_0}\beta^T}{1 - \beta\gamma} \right) \sqrt{\chi_2(\nu \| \pi)\sigma^2 f} \right\}.$$

Proof. Let us start from the bias of the FHN estimator. We proceed as follows, with $t_0 \in \mathbb{N}$:

$$\operatorname*{Bias}_{H_N \sim P^N_{\nu,\rho}} [\widehat{\eta}_{\text{FHN}}(H_N, f)] = \left| \underset{H_N \sim P^N_{\nu,\rho}}{\mathbb{E}} [\widehat{\eta}_{\text{FHN}}(H_N, f)] - \pi_\gamma f \right| \tag{35}$$

$$= (1-\gamma) \left| \sum_{t=0}^{T-1} \gamma^t \underset{X \sim \nu P^t}{\mathbb{E}}[f(X)] - \sum_{t=0}^{+\infty} \gamma^t \underset{X \sim \nu P^t}{\mathbb{E}}[f(X)] \right| \tag{36}$$

$$\leqslant (1-\gamma)\gamma^T \left(\sum_{t=0}^{t_0} \gamma^t \left| \underset{X \sim \nu P^{t+T}}{\mathbb{E}}[f(X)] \right| \right. \tag{37}$$

$$\left. + \sum_{t=t_0}^{+\infty} \gamma^t \left| \underset{X \sim \nu P^{t+T}}{\mathbb{E}}[f(X)] - \underset{X \sim \pi}{\mathbb{E}}[f(X)] \right| + \sum_{t=t_0}^{+\infty} \gamma^t \left| \underset{X \sim \pi}{\mathbb{E}}[f(X)] \right| \right) \tag{38}$$

$$\leqslant (1-\gamma)\gamma^T \left(\frac{1-\gamma^{t_0}}{1-\gamma} + \sqrt{\chi_2(\nu\|\pi)\sigma^2 f} \sum_{t=t_0}^{+\infty} \gamma^t \beta^{t+T} + \frac{\gamma^{t_0}}{1-\gamma} \right) \tag{39}$$

$$\leqslant (1-\gamma)\gamma^T \left(\frac{1}{1-\gamma} + \sqrt{\chi_2(\nu\|\pi)\sigma^2 f} \frac{(\beta\gamma)^{t_0}\beta^T}{1-\beta\gamma} \right) \leqslant \gamma^T, \tag{40}$$

where we obtain (37) from (36) by first collecting γ^T from the summation, summing and subtracting the term $\sum_{t=t_0}^{+\infty} \gamma^t \mathbb{E}_{X \sim \pi}[f(X)]$, and then applying the triangle inequality, we employ Lemma A.3 to write (39), and we let $t_0 \to \infty$ to tighten the bound and obtain the last inequality in (40). Let us now move to the FHC estimator. We consider a similar derivation with $s_0 \in \mathbb{N}$ and $t_0 \in [\![0, T]\!]$:

$$\operatorname*{Bias}_{H_N \sim P^N_{\nu,\rho}} [\widehat{\eta}_{\text{FHC}}(H_N, f)] = \left| \frac{1-\gamma}{1-\gamma^T} \sum_{t=0}^{T-1} \gamma^t \underset{X \sim \nu P^t}{\mathbb{E}}[f(X)] - (1-\gamma) \sum_{t=0}^{+\infty} \gamma^t \underset{X \sim \nu P^t}{\mathbb{E}}[f(X)] \right| \tag{41}$$

$$= (1-\gamma)\gamma^T \left| \sum_{t=0}^{T-1} \frac{\gamma^t}{1-\gamma^T} \underset{X \sim \nu P^t}{\mathbb{E}}[f(X)] - \sum_{t=0}^{+\infty} \gamma^t \underset{X \sim \nu P^{t+T}}{\mathbb{E}}[f(X)] \right| \tag{42}$$

$$\leqslant (1-\gamma)\gamma^T \left(\sum_{t=0}^{t_0-1} \gamma^t + \sum_{t=t_0}^{T-1} \frac{\gamma^t}{1-\gamma^T} \left| \underset{X \sim \nu P^t}{\mathbb{E}}[f(X)] - \underset{X \sim \pi}{\mathbb{E}}[f(X)] \right| \right. \tag{43}$$

$$\left. + \sum_{t=0}^{s_0-1} \gamma^t + \sum_{t=s_0}^{+\infty} \gamma^t \left| \underset{X \sim \nu P^{t+T}}{\mathbb{E}}[f(X)] - \underset{X \sim \pi}{\mathbb{E}}[f(X)] \right| \right) \tag{44}$$

$$\leqslant (1-\gamma)\gamma^T \left(\frac{2-\gamma^{t_0}-\gamma^{s_0}}{1-\gamma} + \left(\frac{1}{1-\gamma^T} \sum_{t=t_0}^{T-1} \gamma^t \beta^t + \sum_{t=s_0}^{+\infty} \gamma^t \beta^{t+T} \right) \sqrt{\chi_2(\nu\|\pi)\sigma^2 f} \right) \tag{45}$$

$$\leqslant (1-\gamma)\gamma^T \left(\frac{2-\gamma^{t_0}-\gamma^{s_0}}{1-\gamma} + \left(\frac{(\beta\gamma)^{t_0}(1-(\beta\gamma)^{T-t_0})}{(1-\gamma^T)(1-\beta\gamma)} + \frac{(\beta\gamma)^{s_0}\beta^T}{1-\beta\gamma} \right) \sqrt{\chi_2(\nu\|\pi)\sigma^2 f} \right), \tag{46}$$

where line (44) follows from triangle inequality and line (45) is obtained by applying Lemma A.3. The result is obtained by making explicit the minimization over t_0 and s_0. $\square$

Corollary 5.2 (FHC Estimator – Bias)**.** *Let* $H_N \sim P^N_{\nu,\rho}$ *with the reset policy* $\rho_t = \delta_{\mathbf{1}\{t \bmod T = 0\}}$*, and let* $f : \mathcal{X} \to [0, 1]$. *The bias of the FHC estimator is upper bounded as:*

- *if* $\beta = 0$:

$$b^{FHC}_{0,\gamma,T} = \frac{1-\gamma}{1-\gamma^T} \gamma^T \min \left\{ \sqrt{\chi_2(\nu\|\pi)\sigma^2 f}, 1-\gamma^T \right\}; \tag{10}$$

- *if* $\beta = 1$:

$$b^{FHC}_{1,\gamma,T} = 2\gamma^T \min \left\{ \sqrt{\chi_2(\nu\|\pi)\sigma^2 f}, 1 \right\}. \tag{11}$$

Proof. We start considering the case $\beta = 0$. Since $T \geqslant 1$, we distinguish between the case in which the optimal value of

t_0 is 0 or grater than 0. Instead, for s_0 it is always convenient to select $s_0 = 0$. If $t_0 = 0$, the bias bound becomes:

$$b_{0,\gamma,T}^{\text{FHC}}|_{t_0=0} = (1-\gamma)\gamma^T \cdot \frac{1}{1-\gamma^T} \sqrt{\chi_2(\nu\|\pi)\sigma^2 f}.$$

If instead, we select $t_0 > 0$, we get:

$$b_{0,\gamma,T}^{\text{FHC}}|_{t_0>0} = (1-\gamma)\gamma^T \cdot \frac{1-\gamma^{t_0}}{1-\gamma} = \gamma^T(1-\gamma^{t_0}),$$

that is minimized by selecting $t_0 = 1$. Thus, putting all together, we obtain the minimum between the two expression, whose value depend on the entity of the term $\sqrt{\chi_2(\nu\|\pi)\sigma^2 f}$. Let us move to the case $\beta = 1$. Now, the bias bound becomes:

$$
\begin{aligned}
b_{1,\gamma,T}^{\text{FHC}} &= (1-\gamma)\gamma^T \left(\frac{2-\gamma^{t_0}-\gamma^{s_0}}{1-\gamma} + \left(\frac{\gamma^{t_0}(1-\gamma^{T-t_0})}{(1-\gamma^T)(1-\gamma)} + \frac{\gamma^{s_0}}{1-\gamma} \right) \sqrt{\chi_2(\nu\|\pi)\sigma^2 f} \right) \\
&= (1-\gamma)\gamma^T \underbrace{\left(\frac{1-\gamma^{t_0}}{1-\gamma} + \frac{\gamma^{t_0}(1-\gamma^{T-t_0})}{(1-\gamma^T)(1-\gamma)} \sqrt{\chi_2(\nu\|\pi)\sigma^2 f} \right)}_{f(t_0)} \\
&\quad + (1-\gamma)\gamma^T \underbrace{\left(\frac{1-\gamma^{s_0}}{1-\gamma} + \frac{\gamma^{s_0}}{1-\gamma} \sqrt{\chi_2(\nu\|\pi)\sigma^2 f} \right)}_{g(s_0)}.
\end{aligned}
$$

We proceed in a separate way for t_0 and s_0, as they can be optimized independently. Let us start with t_0:

$$f(t_0) = \left(\frac{1-\gamma^{t_0}}{1-\gamma} + \frac{\gamma^{t_0}(1-\gamma^{T-t_0})}{(1-\gamma^T)(1-\gamma)} \sqrt{\chi_2(\nu\|\pi)\sigma^2 f} \right).$$

It is simple to see, by renaming $x := \gamma^{t_0}$, that $f(x)$ has no stationary points. Therefore, the optimum must be in the extreme points $t_0 \in \{0, T\}$:

$$f(0) = \frac{1}{1-\gamma} \sqrt{\chi_2(\nu\|\pi)\sigma^2 f},$$

$$f(T) = \frac{1-\gamma^T}{1-\gamma}.$$

Let us now move considering s_0:

$$g(s_0) = \frac{1-\gamma^{s_0}}{1-\gamma} + \frac{\gamma^{s_0}}{1-\gamma} \sqrt{\chi_2(\nu\|\pi)\sigma^2 f}.$$

Similarly to the previous case, function g admits no stationary points and, thus, we consider the extreme values:

$$g(0) = \frac{1}{1-\gamma} \sqrt{\chi_2(\nu\|\pi)\sigma^2 f},$$

$$g(+\infty) = \frac{1}{1-\gamma}.$$

Putting all together, we obtain:

$$
\begin{aligned}
b_{1,\gamma,T}^{\text{FHC}} &= (1-\gamma)\gamma^T \left(\min\{f(0), f(T)\} + \min\{g(0), g(+\infty)\} \right) \\
&= \gamma^T \left(\min\left\{ \sqrt{\chi_2(\nu\|\pi)\sigma^2 f}, 1-\gamma^T \right\} + \left\{ \sqrt{\chi_2(\nu\|\pi)\sigma^2 f}, 1 \right\} \right) \\
&\leqslant 2\gamma^T \min\left\{ \sqrt{\chi_2(\nu\|\pi)\sigma^2 f}, 1 \right\},
\end{aligned}
$$

where the last inequality is to obtain a more interpretable expression. $\qquad\square$

Concentration

Theorem 5.3 (FH Estimators – Concentration). *Let $H_N \sim P_{\nu,\boldsymbol{\rho}}^N$ with the reset policy $\rho_t = \delta_{\mathbf{1}\{t \bmod T=0\}}$, and let $f : \mathcal{X} \to$*

$[0, 1]$. *Let us define for $j_0 \in [\![0, T]\!]$:*

$$c_{\beta,\gamma}(j_0) := \frac{(\beta\gamma)^{j_0} - (\beta\gamma)^T}{1 - \beta\gamma} \sqrt{\chi_2(\nu\|\pi)\sigma^2 f},$$

$$d_{\beta,\gamma}(j_0, \delta) := \sqrt{\frac{8T \left(\log \left(\chi_2(\nu P^{j_0}\|\pi) + 1 \right) + 4 \log \frac{2}{\delta} \right)}{N}}$$

$$\times \sqrt{\frac{(1 - \gamma^{j_0})^2}{(1 - \gamma)^2} + \frac{(1 + \beta)(\gamma^{2j_0} - \gamma^{2T})}{(1 - \beta)(1 - \gamma^2)}}.$$

For every $\delta \in (0, 1)$ with probability at least $1 - \delta$, it holds:

$$
\begin{aligned}
|\widehat{\eta}_{FHN}(H_N, f) - \pi_\gamma f| \leqslant\ & b_{\gamma,T}^{FHN} \\
& + (1 - \gamma) \min_{j_0 \in [\![0, T]\!]} \{ c_{\beta,\gamma}(j_0) + d_{\beta,\gamma}(j_0, \delta) \},
\end{aligned}
\tag{12}
$$

$$
\begin{aligned}
|\widehat{\eta}_{FHC}(H_N, f) - \pi_\gamma f| \leqslant\ & b_{\beta,\gamma,T}^{FHC} \\
& + \frac{1 - \gamma}{1 - \gamma^T} \min_{j_0 \in [\![0, T]\!]} \{ c_{\beta,\gamma}(j_0) + d_{\beta,\gamma}(j_0, \delta) \}.
\end{aligned}
\tag{13}
$$

Proof. We provide a derivation that holds for both the FHN and the FHC estimators. Specifically, we consider a constant $c_{\text{FH}\star}$ with $\star \in \{\text{N,C}\}$, that is differently defined for the FHN and FHC estimators as follows:

$$c_{\text{FH}\star} = \begin{cases} 1 - \gamma & \text{if } \star = \text{N} \\ \frac{1-\gamma}{1-\gamma^T} & \text{if } \star = \text{C} \end{cases}.$$

Let us consider the moment–generating function for $t \in \mathbb{R}$:

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(t \left(c_{\text{FH}\star} \sum_{i=0}^{M-1} \sum_{j=0}^{T-1} \gamma^j f(X_{Ti+j}) - \pi_\gamma f \right) \right) \right] &= \mathbb{E}\left[\prod_{i=0}^{M-1} \exp\left(t \left(c_{\text{FH}\star} \sum_{j=0}^{T-1} \gamma^j f(X_{Ti+j}) - \pi_\gamma f \right) \right) \right] \\
&= \prod_{i=0}^{M-1} \mathbb{E}\left[\exp\left(t \left(c_{\text{FH}\star} \sum_{j=0}^{T-1} \gamma^j f(X_{Ti+j}) - \pi_\gamma f \right) \right) \right] \\
&= \mathbb{E}\left[\exp\left(t \left(c_{\text{FH}\star} \sum_{j=0}^{T-1} \gamma^j f(X_j) - \pi_\gamma f \right) \right) \right]^M,
\end{aligned}
$$

where the last but one equality follows from the fact that each trajectory is independent from the others, since the reset is based on the horizon only, and the last inequality is obtained by observing that the trajectories are identically distributed. Let us now focus on one trajectory only and we highlight a bias term:

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(t \left(c_{\text{FH}\star} \sum_{j=0}^{T-1} \gamma^j f(X_j) - \pi_\gamma f \right) \right) \right] &= \mathbb{E}\left[\exp\left(t c_{\text{FH}\star} \sum_{j=0}^{T-1} \gamma^j (f(X_j) - \nu P^j f) \right) \right] \\
&\qquad \times \exp\left(t \left(c_{\text{FH}\star} \sum_{j=0}^{T-1} \gamma^j \nu P^j f - \pi_\gamma f \right) \right) \\
&\leqslant \mathbb{E}\left[\exp\left(t c_{\text{FH}\star} \sum_{j=0}^{T-1} \gamma^j (f(X_j) - \nu P^j f) \right) \right] \exp\left(t b_{\beta,\gamma,T}^{\text{FH}\star} \right),
\end{aligned}
$$

having observed that the last term corresponds to the actual bias, as bounded in Proposition 5.1. Focusing on the first term, we rename $\widetilde{f}(X_j) := f(X_j) - \nu P^j f$, let $j_0 \in [\![0, T]\!]$ and we apply Hölder's inequality with exponent $q \in [1, +\infty]$:

$$\mathbb{E}\left[\exp\left(t c_{\text{FH}\star} \sum_{j=0}^{T-1} \gamma^j \widetilde{f}(X_j) \right) \right] = \mathbb{E}\left[\exp\left(t c_{\text{FH}\star} \sum_{j=0}^{j_0-1} \gamma^j \widetilde{f}(X_j) \right) \exp\left(t c_{\text{FH}\star} \sum_{j=j_0}^{T-1} \gamma^j \widetilde{f}(X_j) \right) \right]$$

$$\leqslant \mathbb{E}\left[\exp\left(tc_{\text{FH}\star}q\sum_{j=0}^{j_0-1}\gamma^j\widetilde{f}(X_j)\right)\right]^{\frac{1}{q}}\underbrace{\mathbb{E}\left[\exp\left(\frac{tc_{\text{FH}\star}q}{q-1}\sum_{j=j_0+1}^{T-1}\gamma^j\widetilde{f}(X_j)\right)\right]^{\frac{q-1}{q}}}_{\text{(b)}}.$$

$$\underbrace{\phantom{\mathbb{E}\left[\exp\left(tc_{\text{FH}\star}q\sum_{j=0}^{j_0-1}\gamma^j\widetilde{f}(X_j)\right)\right]^{\frac{1}{q}}}}_{\text{(a)}}$$

Let us focus on term (a), we look at the quantity $c_{\text{FH}\star}q\sum_{j=0}^{j_0-1}\gamma^j\widetilde{f}(X_j)$ as a unique random variable whose range is $4c_{\text{FH}\star}q(1-\gamma^{j_0})/(1-\gamma)$ as $|\widetilde{f}(X_j)|\leqslant 1$. Thus, by Höeffding's lemma, having observed that the terms $\widetilde{f}(X_j)$ are zero-mean random variables:

$$\text{(a)} = \mathbb{E}\left[\exp\left(tc_{\text{FH}\star}q\sum_{j=0}^{j_0-1}\gamma^j\widetilde{f}(X_j)\right)\right]^{\frac{1}{q}} \leqslant \exp\left(\frac{2t^2c_{\text{FH}\star}^2 q(1-\gamma^{j_0})^2}{(1-\gamma)^2}\right).$$

Let us move to term (b). Here, we need to highlight a further bias term:

$$\text{(b)} = \mathbb{E}\left[\exp\left(\frac{tc_{\text{FH}\star}q}{q-1}\sum_{j=j_0+1}^{T-1}\gamma^j\widetilde{f}(X_j)\right)\right]^{\frac{q-1}{q}}$$

$$= \mathbb{E}\left[\exp\left(\frac{tc_{\text{FH}\star}q}{q-1}\sum_{j=j_0+1}^{T-1}\gamma^j(f(X_j)-\nu P^j f)\right)\right]^{\frac{q-1}{q}}$$

$$= \underbrace{\mathbb{E}\left[\exp\left(\frac{tc_{\text{FH}\star}q}{q-1}\sum_{j=j_0+1}^{T-1}\gamma^j(f(X_j)-\pi f)\right)\right]^{\frac{q-1}{q}}}_{\text{(c)}}\exp\left(tc_{\text{FH}\star}\sum_{j=j_0+1}^{T-1}\gamma^j(\pi f - \nu P^j f)\right).$$

Now, we focus on term (c) and proceed with a change of measure followed by an application of Hölder's inequality with exponent $r\in[1,+\infty]$:

$$\text{(c)} = \mathbb{E}\left[\exp\left(\frac{tc_{\text{FH}\star}q}{q-1}\sum_{j=j_0+1}^{T-1}\gamma^j(f(X_j)-\pi f)\right)\right]^{\frac{q-1}{q}}$$

$$= \mathbb{E}_{\pi}\left[\frac{\nu P^{j_0}(X_{j_0})}{\pi(X_{j_0})}\exp\left(\frac{tc_{\text{FH}\star}q}{q-1}\sum_{j=j_0+1}^{T-1}\gamma^j(f(X_j)-\pi f)\right)\right]^{\frac{q-1}{q}}$$

$$\leqslant \mathbb{E}_{\pi}\left[\left(\frac{\nu P^{j_0}(X_{j_0})}{\pi(X_{j_0})}\right)^r\right]^{\frac{q-1}{rq}}\underbrace{\mathbb{E}_{\pi}\left[\exp\left(\frac{tc_{\text{FH}\star}qr}{(q-1)(r-1)}\sum_{j=j_0+1}^{T-1}\gamma^j(f(X_j)-\pi f)\right)\right]^{\frac{(q-1)(r-1)}{qr}}}_{\text{(d)}}.$$

Then, we consider term (e) and apply Theorem 1 of Fan et al. (2021) to bound the moment generating function, recalling that $1-\beta$ is the absolute spectral gap:

$$\text{(d)} = \mathbb{E}_{\pi}\left[\exp\left(\frac{tc_{\text{FH}\star}qr}{(q-1)(r-1)}\sum_{j=j_0+1}^{T-1}\gamma^j(f(X_j)-\pi f)\right)\right]^{\frac{(q-1)(r-1)}{qr}}$$

$$\leqslant \exp\left(\frac{2t^2c_{\text{FH}\star}^2 qr}{(q-1)(r-1)}\cdot\frac{1+\beta}{1-\beta}\cdot\frac{\gamma^{2j_0}-\gamma^{2T}}{1-\gamma^2}\right).$$

Concerning the second bias term, we can provide a bound by exploiting Lemma A.3:

$$\sum_{j=j_0+1}^{T-1}\gamma^j(\pi f - \nu P^j f) \leqslant \sum_{j=j_0+1}^{T-1}(\beta\gamma)^j\sqrt{\chi_2(\nu\|\pi)\sigma^2 f} = \frac{(\beta\gamma)^{j_0}-(\beta\gamma)^T}{1-\beta\gamma}\sqrt{\chi_2(\nu\|\pi)\sigma^2 f} := c_{\beta,\gamma}(j_0).$$

Alberto Maria Metelli, Mirco Mutti, Marcello Restelli

Putting all together and by minimizing over t, we have, for $\star \in \{N,C\}$:

$$\Pr_{H_N \sim P_{\nu,\boldsymbol{\rho}}^N} \left(|\widehat{\eta}_{\text{FH}\star}(H_N, f) - \pi_\gamma f| > \epsilon\right) \leqslant 2 \left\|\frac{\nu P^{j_0}}{\pi}\right\|_{r,\pi}^{\frac{q-1}{q}}$$

$$\times \min_{t \geqslant 0} \exp\left(-t\left(\epsilon - b_{\beta,\gamma,T}^{\text{FH}\star} - c_{\text{FH}\star} c_{\beta,\gamma}(j_0)\right) \right.$$

$$\left. + 2t^2 c_{\text{FH}\star}^2 q \left(\frac{(1-\gamma^{j_0})^2}{(1-\gamma)^2} + \frac{r}{(q-1)(r-1)} \cdot \frac{1+\beta}{1-\beta} \cdot \frac{\gamma^{2j_0}-\gamma^{2T}}{1-\gamma^2}\right) \right)^M$$

$$= 2\left\|\frac{\nu P^{j_0}}{\pi}\right\|_{r,\pi}^{\frac{q-1}{q}} \exp\left(-\frac{\left(\epsilon - b_{\beta,\gamma,T}^{\text{FH}\star} - c_{\text{FH}\star} c_{\beta,\gamma}(j_0)\right)^2 M}{8c_{\text{FH}\star}^2 q\left(\frac{(1-\gamma^{j_0})^2}{(1-\gamma)^2} + \frac{r}{(q-1)(r-1)} \cdot \frac{1+\beta}{1-\beta} \cdot \frac{\gamma^{2j_0}-\gamma^{2T}}{1-\gamma^2}\right)} \right).$$

By solving for ϵ, and minimizing over the free parameters q, r, and j_0, we obtain that with probability at least $1-\delta$ it holds that:

$$|\widehat{\eta}_{\text{FH}\star}(H_N, f) - \pi_\gamma f| \leqslant b_{\beta,\gamma,T}^{\text{FH}\star} + c_{\text{FH}\star} \min_{j_0 \in [\![0,T]\!], q,r \geqslant 1} \left\{ c_{\beta,\gamma}(j_0) \right.$$

$$\left. + \sqrt{\frac{8q}{M}\left(\frac{(1-\gamma^{j_0})^2}{(1-\gamma)^2} + \frac{r}{(q-1)(r-1)} \cdot \frac{1+\beta}{1-\beta} \cdot \frac{\gamma^{2j_0}-\gamma^{2T}}{1-\gamma^2}\right)} \right\}$$

$$\times \sqrt{\left(\log\frac{2}{\delta} + \log\left\|\frac{\nu P^{j_0}}{\pi}\right\|_{r,\pi}^{\frac{q-1}{q}}\right)}.$$

To obtain the theorem statement, we set $q = r = 2$, to get:

$$|\widehat{\eta}_{\text{FH}\star}(H_N, f) - \pi_\gamma f| \leqslant b_{\beta,\gamma,T}^{\text{FH}\star} + c_{\text{FH}\star} \min_{j_0 \in [\![0,T]\!]} c_{\beta,\gamma}(j_0) + \sqrt{\frac{16}{M}\left(\frac{(1-\gamma^{j_0})^2}{(1-\gamma)^2} + 2 \cdot \frac{1+\beta}{1-\beta} \cdot \frac{\gamma^{2j_0}-\gamma^{2T}}{1-\gamma^2}\right)}$$

$$\times \sqrt{\left(\log\frac{2}{\delta} + \frac{1}{4}\log\left\|\frac{\nu P^{j_0}}{\pi}\right\|_{2,\pi}^2\right)}$$

$$\leqslant b_{\beta,\gamma,T}^{\text{FH}\star} + c_{\text{FH}\star} \min_{j_0 \in [\![0,T]\!]} c_{\beta,\gamma}(j_0) + \sqrt{\left(\frac{(1-\gamma^{j_0})^2}{(1-\gamma)^2} + \frac{1+\beta}{1-\beta} \cdot \frac{\gamma^{2j_0}-\gamma^{2T}}{1-\gamma^2}\right)}$$

$$\times \sqrt{\frac{32}{M}\left(\log\frac{2}{\delta} + \frac{1}{4}\log\left(\chi_2(\nu P^{j_0}\|\pi)+1\right)\right)},$$

the statement is obtained by observing that $M = T/N$. $\qquad\square$

Corollary 5.4 (FH Estimators – Concentration). *Let $H_N \sim P_{\nu,\boldsymbol{\rho}}^N$ with the reset policy $\rho_t = \delta_{\mathbf{1}\{t \bmod T=0\}}$, and let $f : \mathcal{X} \to [0,1]$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that:[7]*

- *if $\beta = 0$:*

$$|\widehat{\eta}_{FHN}(H_N, f) - \pi_\gamma f| \leqslant O\left(\gamma^T + \sqrt{\frac{T(1-\gamma)(1-\gamma^T)\log\frac{2}{\delta}}{N}}\right), \tag{14}$$

$$|\widehat{\eta}_{FHC}(H_N, f) - \pi_\gamma f| \leqslant O\left((1-\gamma)\gamma^T + \sqrt{\frac{T(1-\gamma)\log\frac{2}{\delta}}{N(1-\gamma^T)}}\right); \tag{15}$$

- *if $\beta = 1$:*

$$|\widehat{\eta}_{FHN}(H_N, f) - \pi_\gamma f| \leqslant O\left(\gamma^T + (1-\gamma^T)\sqrt{\frac{T\log\frac{2}{\delta}}{N}}\right), \tag{16}$$

[7]For interpretability reasons, we ignore the dependence on $\chi_2(\nu\|\pi)\sigma^2 f$.

$$|\widehat{\eta}_{FHC}(H_N, f) - \pi_\gamma f| \leqslant O\left(\gamma^T + \sqrt{\frac{T \log \frac{2}{\delta}}{N}}\right). \tag{17}$$

Proof. We start with the case $\beta = 1$. From Theorem 5.3, we immediately observe that we need to select $j_0 = T$, otherwise the concentration bound degenerates to infinity. Moreover, we make use of the bias bounds of Corollary 5.4. Thus, ignoring the term $\chi_2(\nu \| \pi)\sigma^2 f$, we obtain the O expression provided in the statement of the corollary. Concerning the case $\beta = 0$, instead, we need some additional care. First of all, we observe that $c_{0,\gamma}(j_0) = 0$ for every $j_0 > 0$, and $c_{0,\gamma}(0) = \sqrt{\chi_2(\nu \| \pi)\sigma^2 f}$. Thus, we can ignore this term. Then, it is simple to verify that the expression $d_{0,\gamma}(j_0)$, ignoring the dependence on $\chi_2(\nu \| \pi)\sigma^2 f$ again, is given by:

$$d_{0,\gamma}(j_0) = O\left(\sqrt{\frac{(1 - \gamma^{j_0})^2}{(1 - \gamma)^2} + \frac{\gamma^{2j_0} - \gamma^{2T}}{1 - \gamma^2}} \sqrt{\frac{T \log \frac{2}{\delta}}{N}}\right).$$

By vanishing the derivative, we obtain the value of j_0 that is minimizing the expression, i.e., $j_0 = \frac{\log((1+\gamma)/2)}{\log \gamma}$. This quantity is in the interval $[0, 1/2]$ varying $\gamma \in [0, 1]$. Consequently, as j_0 must be integer, we select $j_0 = 0$, to get the expression shown in the corollary statement. □

A.2.2 Adaptive-Horizon Estimation Algorithms

Computational Analysis

Theorem 5.5 (AH Estimators – Complexity). *Let $H_N \sim P_{\nu,\rho}^N$ with the reset policy $\rho_t^{AHR}(H_t, X_t) = \text{Ber}(1 - \gamma)$. Then, the number of trajectories M is distributed such that $M - 1 \sim \text{Bin}(N - 1, 1 - \gamma)$. Furthermore, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$, the time complexity is bounded as:*

$$\max_{i \in [\![M]\!]} T_i \leqslant O\left(\min\left\{N, \frac{\log\left(N^2/\delta\right)}{1 - \gamma}\right\}\right).$$

Proof. To characterize the number of trajectories, we consider the sampling process, in which, at every step t, we sample independently a Bernoulli random variable to decide whether to reset:

$$Y_t \sim \rho_t^{\text{AHR}}(H_t, X_t) = \text{Ber}(1 - \gamma).$$

We have already observed that the number of trajectories can be computed as $M = 1 + \sum_{t=0}^{N-1} Y_i$. Consequently, we have that $M - 1$ is the sum of N independent Bernoulli random variables, being a binomial random variable $\text{Bin}(N - 1, 1 - \gamma)$. From the properties of the binomial random variable, we have that $\mathbb{E}[M - 1] = (N - 1)(1 - \gamma)$.

To analyze the time complexity, we need to characterize the distribution of the maximum length among the trajectories, i.e., $T_{\max} = \max_{i \in [\![M]\!]} T_i$. Each T_i can be looked as derived from a geometric distribution as $T_i - 1 \sim \text{Geo}(1 - \gamma)$. Unfortunately, these random variables are dependent (but identically distributed) since the process stops as soon as the have run out of budget. To this end, we will proceed as follows, being $k \in \mathbb{N}$:

$$\begin{aligned}
\Pr\left(\max_{i \in [\![M]\!]} T_{i-1} > k\right) &= \mathbb{E}\left[\mathbf{1}\left\{\max_{i \in [\![M]\!]} T_{i-1} > k\right\}\right] \\
&= \mathbb{E}\left[\sum_{m=1}^{N} \mathbf{1}\{M = m\} \mathbf{1}\left\{\max_{i \in [\![M]\!]} T_{i-1} > k\right\}\right] \\
&\leqslant \sum_{m=1}^{N} \mathbb{E}\left[\mathbf{1}\left\{\max_{i \in [\![m]\!]} T_{i-1} > k\right\}\right] \\
&= \sum_{m=1}^{N} \Pr\left(\max_{i \in [\![m]\!]} T_{i-1} > k\right).
\end{aligned}$$

Now, we consider one term at a time and perform a union bound:

$$\Pr\left(\max_{i\in[\![m]\!]} T_{i-1} > k\right) = \Pr\left(\bigvee_{i\in[\![m]\!]} T_{i-1} > k\right)$$
$$\leqslant \sum_{i\in[\![m]\!]} \Pr\left(T_{i-1} > k\right)$$
$$= m\Pr\left(T_0 > k\right),$$

where the last equality follows from the fact that the random variables T_i are identically distributed. Since $T_0 - 1$ is a geometric distributions, we have that $\Pr(T_0 \geqslant k) = \gamma^{k-1}$. Thus, putting all together, we obtain:

$$\sum_{m=1}^{N} \Pr\left(\max_{i\in[\![m]\!]} T_{i-1} > k\right) = \sum_{m=1}^{N} m\gamma^{k-1} = \frac{N(N+1)}{2}\gamma^{k-1}.$$

Solving to obtain k, we have that with probability at least $1 - \delta$ it holds that:

$$k \leqslant 1 + \frac{\log\frac{N(N+1)}{2\delta}}{\log\frac{1}{\gamma}} \leqslant 1 + \frac{\log\frac{2N^2}{2\delta}}{1-\gamma},$$

having observed that $N + 1 \leqslant N$ and $\frac{1}{\log\frac{1}{\gamma}} \leqslant \frac{1}{1-\gamma}$. By taking the minimum with the number of samples N, we get the result. $\square$

Statistical Analysis

Theorem 5.6 (OS Estimator – Concentration). *Let $H_N \sim P^N_{\nu,\rho}$ with the reset policy $\rho_t^{AHR}(H_t, X_t) = \mathrm{Ber}(1-\gamma)$, and let $f : \mathcal{X} \to [0,1]$. For every $\delta \in (0,1)$, with probability at least $1 - \delta$, it holds that:*

$$|\widehat{\eta}_{OS}(H_N, f) - \pi_\gamma f| \leqslant \sqrt{\frac{2\log\frac{8}{\delta}}{N(1-\gamma)}}.$$

Proof. Suppose that the number of trajectories $M - 1 = m - 1$ is fixed. In this case, we can apply Höeffding's inequality to the estimator:[8]

$$\Pr_{H_N \sim P^N_{\nu,\rho^{AHR}}} (|\widehat{\eta}_{OS}(H_N, f) - \pi_\gamma f| > \epsilon | M = m) \leqslant 2\exp\left(-\frac{\epsilon^2(m-1)}{2}\right).$$

Now, we take the expectation w.r.t. to the distribution of M that is a binomial distribution:

$$\Pr_{H_N \sim P^N_{\nu,\rho^{AHR}}} (|\widehat{\eta}_{OS}(H_N, f) - \pi_\gamma f| > \epsilon) = \mathbb{E}_{m \sim \mathrm{Bin}(N, 1-\gamma)}\left[\Pr_{H_N \sim P^N_{\nu,\rho^{AHR}}} (|\widehat{\eta}_{OS}(H_N, f) - \pi_\gamma f| > \epsilon | M = m)\right]$$
$$\leqslant \mathbb{E}_{m \sim \mathrm{Bin}(N, 1-\gamma)}\left[2\exp\left(-\frac{\epsilon^2 m}{2}\right)\right]$$
$$= 2\sum_{m=0}^{N} \binom{N}{m}(1-\gamma)^m \gamma^{N-m} \exp\left(-\frac{\epsilon^2 m}{2}\right)$$
$$= 2\left(\gamma + (1-\gamma)\exp\left(-\frac{\epsilon^2}{2}\right)\right)^N.$$

We now provide a looser but more interpretable bound. To this end, we consider the derivation, holding for $\epsilon \in [0,1]$ (since

---

[8]Note that conditioning to $M = m$ is allowed as the decision to reset is independent on the values of $f(X_t)$ but depends on an independent trial $Y_t \sim \mathrm{Ber}(1-\gamma)$ at each step.

$f(x) \in [0, 1]$ for all $x \in \mathcal{X}$):

$$\frac{\gamma + (1 - \gamma) \exp\left(-\frac{\epsilon^2}{2}\right)}{\exp\left(-\frac{\epsilon^2(1-\gamma)}{2}\right)} = \gamma \exp\left(\frac{\epsilon^2(1-\gamma)}{2}\right) + (1 - \gamma) \exp\left(-\frac{\epsilon^2 \gamma}{2}\right)$$

$$\begin{cases} \leqslant 1 + e \\ \geqslant \gamma + (1 - \gamma) \exp(-\gamma) \geqslant \exp(-\gamma) \geqslant e^{-1} \end{cases}$$

Thus, we have that with probability at least $1 - \delta$ it holds that:

$$|\widehat{\eta}_{\mathrm{OS}}(H_N, f) - \pi_\gamma f| \leqslant \sqrt{\frac{2 \log \frac{2(1+e)}{\delta}}{N(1-\gamma)}}.$$

The result is obtained by observing that $2(1 + e) < 8$. $\qquad\square$

Theorem 5.7 (AS Estimator – Concentration). *Let $H_N \sim P_{\nu,\rho}^N$ with the reset policy $\rho_t^{AHR}(H_t, X_t) = \mathrm{Ber}(1 - \gamma)$, and let $f : \mathcal{X} \to [0, 1]$. For every $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that:*

$$|\widehat{\eta}_{AS}(H_N, f) - \pi_\gamma f| \leqslant \sqrt{\frac{8 \log \frac{2}{\delta} + 4 \log(\chi_2(\nu \| \pi_\gamma) + 1)}{N(1 - \beta\gamma)}}.$$

Proof. We start by working on the moment-generating function. Let $t \in \mathbb{R}$. Let us consider $i_0 \in [\![0, N]\!]$:

$$\mathbb{E}\left[\exp\left(t \sum_{i=0}^{N-1} (f(X_i) - \pi_\gamma f)\right)\right] = \mathbb{E}\left[\exp\left(t \sum_{i=0}^{i_0-1} (f(X_i) - \pi_\gamma f)\right) \exp\left(t \sum_{i=i_0}^{N-1} (f(X_i) - \pi_\gamma f)\right)\right]$$

$$\leqslant \exp(i_0 t) \mathbb{E}\left[\exp\left(t \sum_{i=i_0}^{N-1} (f(X_i) - \pi_\gamma f)\right)\right],$$

where we exploited the inequality $|f(X_i) - \pi_\gamma f| \leqslant 1$. We now move to bound the second term, by exploiting a change of measure argument and Hölder's inequality with $q \in [1, +\infty]$:

$$\mathbb{E}\left[\exp\left(t \sum_{i=i_0}^{N-1} (f(X_i) - \pi_\gamma f)\right)\right] = \mathbb{E}_{\pi_\gamma}\left[\frac{\nu P_\gamma^{i_0}(X_{i_0})}{\pi_\gamma(X_{i_0})} \exp\left(t \sum_{i=i_0}^{N-1} (f(X_i) - \pi_\gamma f)\right)\right]$$

$$\leqslant \mathbb{E}_{\pi_\gamma}\left[\left(\frac{\nu P_\gamma^{i_0}(X_{i_0})}{\pi_\gamma(X_{i_0})}\right)^q\right]^{\frac{1}{q}} \mathbb{E}_{\pi_\gamma}\left[\exp\left(\frac{tq}{q-1} \sum_{i=i_0}^{N-1} (f(X_i) - \pi_\gamma f)\right)\right]^{\frac{q-1}{q}}$$

$$= \left\|\frac{\nu P_\gamma^{i_0}}{\pi_\gamma}\right\|_{\pi_\gamma, q} \mathbb{E}_{\pi_\gamma}\left[\exp\left(\frac{tq}{q-1} \sum_{i=i_0}^{N-1} (f(X_i) - \pi_\gamma f)\right)\right]^{\frac{q-1}{q}}$$

Now, we exploit Lemma A.4 to derive that the absolute spectral gap of P_γ is $1 - \beta\gamma$, being $\beta\gamma$ the second eigenvalue of operator P_γ. To bound the expectation in the previous equation, we exploit Theorem 1 of Fan et al. (2021):

$$\mathbb{E}_{\pi_\gamma}\left[\exp\left(\frac{tq}{q-1} \sum_{i=i_0}^{N-1} (f(X_i) - \pi_\gamma f)\right)\right]^{\frac{q-1}{q}} \leqslant \exp\left(\frac{2t^2 q}{q-1} \cdot (N - i_0) \cdot \frac{1 + \beta\gamma}{1 - \beta\gamma}\right).$$

We can now proceed to bound the probability, by minimizing over $t \geqslant 0$:

$$\Pr_{H_N \sim P_{\nu,\rho}^N}(|\widehat{\eta}_{\mathrm{AS}}(H_N, f) - \pi_\gamma f| > \epsilon) \leqslant 2 \left\|\frac{\nu P_\gamma^{i_0}}{\pi_\gamma}\right\|_{\pi_\gamma, q} \min_{t \geqslant 0} \exp\left(-t(\epsilon N - i_0) + \frac{2t^2 q}{q-1} \cdot (N - i_0) \cdot \frac{1 + \beta\gamma}{1 - \beta\gamma}\right)$$

$$= 2 \left\|\frac{\nu P_\gamma^{i_0}}{\pi_\gamma}\right\|_{\pi_\gamma, q} \exp\left(-\frac{(\epsilon N - i_0)^2}{\frac{2q}{q-1} \cdot (N - i_0) \cdot \frac{1+\beta\gamma}{1-\beta\gamma}}\right)$$

$$= 2 \left\| \frac{\nu P_\gamma^{i_0}}{\pi_\gamma} \right\|_{\pi_\gamma, q} \exp\left(-(\epsilon N - i_0)^2 \cdot \frac{q-1}{2q(N-i_0)} \cdot \frac{1 - \beta\gamma}{1 + \beta\gamma} \right).$$

By solving for ϵ, and minimizing over the free parameters q and i_0, we obtain that with probability at least $1 - \delta$ it holds that:

$$|\widehat{\eta}_{\mathrm{AS}}(H_N, f) - \pi_\gamma f| \leqslant \min_{i_0 \in [\![0, N]\!], q \geqslant 1} \frac{i_0}{N} + \sqrt{\frac{2q(1 - i_0/N)}{N(q-1)} \cdot \frac{1 + \beta\gamma}{1 - \beta\gamma} \left(\log \frac{2}{\delta} + \log \left\| \frac{\nu P_\gamma^{i_0}}{\pi_\gamma} \right\|_{\pi_\gamma, q} \right)}.$$

Since the optimization is non-trivial, the result shown in the statement of the theorem is obtained by setting $q = 2$ and $i_0 = 0$, observing that $\left\| \frac{\nu}{\pi_\gamma} \right\|_2^2 = \chi_2(\nu \| \pi_\gamma) + 1$ and bounding $1 + \beta\gamma \leqslant 2$. □

Proposition A.2 (AS Estimator - Bias). *Let $H_N \sim P_{\nu, \rho}^N$ with the reset policy $\rho_t^{AHR}(H_t, X_t) = \mathrm{Ber}(1 - \gamma)$, and let $f : \mathcal{X} \to [0, 1]$. Then, it holds that:*

$$\mathop{\mathbb{B}\mathrm{ias}}_{H_N \sim P_{\nu, \rho}^N} [\widehat{\eta}_{\mathrm{AS}}(H_N, f)] \leqslant \frac{1 - (\beta\gamma)^N}{N(1 - \beta\gamma)} \sqrt{\chi_2(\nu \| \pi_\gamma) \sigma_\gamma^2 f}.$$

Proof. Let us consider the following derivation:

$$\mathop{\mathbb{B}\mathrm{ias}}_{H_N \sim P_{\nu, \rho}^N} [\widehat{\eta}_{\mathrm{AS}}(H_N, f)] = \left| \frac{1}{N} \sum_{t=1}^N \mathop{\mathbb{E}}_{H_N \sim P_{\nu, \rho}^N} [\widehat{\eta}_{\mathrm{AS}}(H_N, f)] - \pi_\gamma f \right|$$

$$= \left| \frac{1}{N} \sum_{t=1}^N \mathop{\mathbb{E}}_{X \sim \nu P_\gamma^t} \mathbb{E}[f(X)] - \pi_\gamma f \right|$$

$$\leqslant \sqrt{\chi_2(\nu \| \pi_\gamma) \sigma_\gamma^2 f} \frac{1}{N} \sum_{t=0}^{N-1} (\beta\gamma)^t \tag{47}$$

$$= \sqrt{\chi_2(\nu \| \pi_\gamma) \sigma_\gamma^2 f} \frac{1 - (\beta\gamma)^N}{N(1 - \beta\gamma)}.$$

where line (47) follows from Lemma A.3 an d recalling that the absolute spectral gap of P_γ is $\beta\gamma$. □

A.2.3 About the Optimal Horizon T

In this appendix, we elaborate on the choice of the horizon T for the FH estimators in a way that it is independent on the mixing properties of the Markov chain. To this end, we consider the simplified expressions of the concentration bounds of Corollary 5.4. Let us define:

$$T_\gamma^* := \frac{\log \sqrt{N}}{\log \frac{1}{\gamma}}.$$

Finite Horizon Corrected Estimator For the FHC estimator we show that the choice of $T = T_\gamma^*$ makes the concentration rate nearly minimax optimal. Indeed, for $\beta = 0$, we have:

$$|\widehat{\eta}_{\mathrm{FHC}}(H_N, f) - \pi_\gamma f| \leqslant O\left((1 - \gamma)\gamma^{T_\gamma^*} + \sqrt{\frac{T_\gamma^*(1 - \gamma) \log \frac{2}{\delta}}{N(1 - \gamma^{T_\gamma^*})}} \right)$$

$$= O\left(\frac{1 - \gamma}{\sqrt{N}} + \sqrt{\frac{\log \sqrt{N}(1 - \gamma) \log \frac{2}{\delta}}{N \log \frac{1}{\gamma} \left(1 - \frac{1}{\sqrt{N}} \right)}} \right)$$

$$\leqslant O\left(\sqrt{\frac{\log N \log \frac{2}{\delta}}{N}} \right)$$

$$= \widetilde{O}\left(\frac{1}{\sqrt{N}}\right).$$

having observed that $(1-\gamma)/\log(1/\gamma) \leqslant 1$ and whenever $N \geqslant 2$. This concentration rate is indeed matching, in $\widetilde{O}$ sense, the minimax rate. We consider now $\beta = 1$. A similar derivation applies:

$$|\widehat{\eta}_{\text{FHC}}(H_N, f) - \pi_\gamma f| \leqslant O\left(\gamma^{T_\gamma^*} + \sqrt{\frac{T_\gamma^* \log\frac{2}{\delta}}{N}}\right)$$

$$\leqslant O\left(\frac{1}{\sqrt{N}} + \sqrt{\frac{\log\sqrt{N}\log\frac{2}{\delta}}{N\log\frac{1}{\gamma}}}\right)$$

$$\leqslant O\left(\sqrt{\frac{\log N\log\frac{2}{\delta}}{2N(1-\gamma)}}\right)$$

$$= \widetilde{O}\left(\frac{1}{\sqrt{N(1-\gamma)}}\right),$$

having bounded $1/\log(1/\gamma) \leqslant 1/(1-\gamma)$. This rate matches as well, in the $\widetilde{O}$ sense, the minimax concentration rate.

Finite Horizon Non-Corrected Estimator The choice of $T = T_\gamma^*$ happens to make also the concentration rate of the FHN estimator nearly minimax optimal (in the $\widetilde{O}$ sense) for $\beta = 1$. Indeed, we have:

$$|\widehat{\eta}_{\text{FHN}}(H_N, f) - \pi_\gamma f| \leqslant O\left(\gamma^{T_\gamma^*} + \sqrt{\frac{T_\gamma^* \log\frac{2}{\delta}}{N}}\right)$$

$$\leqslant O\left(\frac{1}{\sqrt{N}} + \sqrt{\frac{\log\sqrt{N}\log\frac{2}{\delta}}{N\log\frac{1}{\gamma}}}\right)$$

$$\leqslant O\left(\sqrt{\frac{\log N\log\frac{2}{\delta}}{N(1-\gamma)}}\right)$$

$$= \widetilde{O}\left(\frac{1}{\sqrt{N(1-\gamma)}}\right).$$

For the case $\beta = 0$, we have:

$$|\widehat{\eta}_{\text{FHN}}(H_N, f) - \pi_\gamma f| \leqslant O\left(\gamma^{T_\gamma^*} + \sqrt{\frac{T_\gamma^*(1-\gamma)(1-\gamma^{T_\gamma^*})\log\frac{2}{\delta}}{N}}\right)$$

$$\leqslant O\left(\frac{1}{\sqrt{N}} + \sqrt{\frac{(1-\gamma)\log\sqrt{N}\log\frac{2}{\delta}}{N\log\frac{1}{\gamma}}}\right)$$

$$= \widetilde{O}\left(\frac{1}{\sqrt{N}}\right).$$

A.2.4 About Minimax Optimality of FH Estimators for generic $\beta \in (0, 1)$

In this appendix, we provide further elaboration about the possible minimax optimality of the FH estimators for a generic value of $\beta \in (0, 1)$. Specifically, we show that, according to our analysis, there exists a regime of large values of β, i.e., $\beta \in (\overline{\beta}, 1)$, for which our bound of Theorem 5.3 cannot match the minimax lower bound. To this end, we consider a

simplified version of the bound of Theorem 5.3, that disregards the bias term $c_{\beta,\gamma}(j_0)$ and simply focus on the term:

$$f(j_0) := \frac{(1 - \gamma^{j_0})^2}{(1 - \gamma)^2} + \frac{1 + \beta}{1 - \beta} \cdot \frac{\gamma^{2j_0} - \gamma^{2T}}{1 - \gamma^2}.$$

Let us minimize this term over $j_0 \in [\![0, T]\!]$. We can proceed by vanishing the derivative of f in j_0. It is simple to understand (e.g., by performing the substitution $x = \gamma^{j_0}$, obtaining a quadratic function in x) that the only stationary point j_0^* is the global minimum. However, it might be the case that such point is larger than T. In this case, we need to clip j_0^* to T. Thus, we have:

$$j_0^* = \begin{cases} \frac{\log \frac{(1 - \beta)(1 + \gamma)}{2(1 - \beta\gamma)}}{\log \gamma} & \text{if } \beta \leqslant \frac{1 + \gamma - 2\gamma^T}{1 + \gamma - 2\gamma^{T+1}} \\ T & \text{otherwise} \end{cases}.$$

We let $\overline{\beta} = \frac{1 + \gamma - 2\gamma^T}{1 + \gamma - 2\gamma^{T+1}}$ and show that for $\beta \in (\overline{\beta}, 1)$ the FH estimators are not minimax optimal. To show this we consider first the FHN estimator, that leads to a bound of the form:

$$\gamma^T + (1 - \gamma^T)\sqrt{\frac{T}{N}}.$$

First of all, we observe that the value of T minimizing the previous expression must be sublinear in N, because the second addendum will not shrink as $N \to +\infty$ otherwise. Thus, w.l.o.g., we consider the case $T/N \leqslant \frac{1}{4}$. We have:

$$\gamma^T + (1 - \gamma^T)\sqrt{\frac{T}{N}} \geqslant \frac{\gamma^T}{2} + \frac{1}{2}\sqrt{\frac{T}{N}}.$$

By applying Lemma A.5, we obtain that the minimum value of this function over T, for $\gamma \geqslant 0.3$, is given by:

$$\frac{1}{\sqrt{24N(1 - \gamma)}}.$$

Thus, we conclude that FHN cannot match the minimax lower bound. A similar derivation can be set for the FHC estimator which leads to the bound ($0.3 \leqslant \gamma < 1$):

$$(1 - \gamma)\gamma^T + \sqrt{\frac{T}{N}} \geqslant \frac{1}{\sqrt{6N(1 - \gamma)}}.$$

A.3 Technical Lemmas

Lemma A.3. *Let π be an invariant measure of a Markov chain P with spectral gap $1 - \beta \in (0, 1]$ and initial-state distribution ν. For any bounded measurable function $f \in \mathscr{B}(\mathcal{X})$, it holds that:*

$$\left| \nu P^t f - \pi f \right| \leqslant \sqrt{\chi_2(\nu\|\pi)}\beta^t\sqrt{\sigma^2 f},$$

where $\sigma^2 f = \pi(f - \mathbf{1}\pi f)^2$ is the variance of f under π.

Proof. We exploit the fact that $\nu\Pi = \pi\Pi = \pi$ and that $\pi P^t = \pi$ to derive the following identity for any $c \in \mathbb{R}$:

$$\nu P^t f - \pi f = \nu P^t f - \nu\Pi f = (\nu - \pi)(P^t - \Pi)f = (\nu - \pi)(P^t - \Pi)(f - c).$$

Then, we start from the right-hand side of the identity to write

$$\int_{\mathcal{X}} (\nu(\mathrm{d}x) - \pi(\mathrm{d}x)) \int_{\mathcal{X}} (P^t(\mathrm{d}y|x) - \pi(\mathrm{d}y))(f(y) - c)$$

$$= \int_{\mathcal{X}} \pi(\mathrm{d}x)\frac{\nu(\mathrm{d}x) - \pi(\mathrm{d}x)}{\pi(\mathrm{d}x)} \int_{\mathcal{X}} (P^t(\mathrm{d}y|x) - \pi(\mathrm{d}y))(f(y) - c)$$

$$\leqslant \left(\int_{\mathcal{X}} \pi(\mathrm{d}x) \left(\frac{|\nu(\mathrm{d}x) - \pi(\mathrm{d}x)|}{\pi(\mathrm{d}x)} \right)^p \right)^{\frac{1}{p}} \left(\int_{\mathcal{X}} \pi(\mathrm{d}x) \left| \int_{\mathcal{X}} (P^t(\mathrm{d}y|x) - \pi(\mathrm{d}y))(f(y) - c) \right|^q \right)^{\frac{1}{q}}$$

$$= \left\| \frac{\nu}{\pi} - 1 \right\|_{\pi,p} \left\| (P^t - \Pi)(f - c) \right\|_{\pi,q}$$

$$\leqslant \left\| \frac{\nu}{\pi} - 1 \right\|_{\pi,p} \left\| P^t - \Pi \right\|_{\pi,q \to q} \left\| f - c \right\|_{\pi,q}$$

by applying the Hölder's inequality with $p^{-1} + q^{-1} = 1$, and by taking the supremum over $g = f - c$ to obtain the last line. Finally, we take $p = q = 2$ and $c = \pi f$ to prove the result. $\qquad \square$

Lemma A.4. *Let $P : \mathcal{X} \to \mathscr{P}(\mathcal{X})$ be a Markov kernel, $\nu \in \mathscr{P}(\mathcal{X})$ be the initial state distribution, and $\gamma \in [0,1]$ be the discount factor. Let $P_\gamma = (1 - \gamma)\mathbf{1}\nu + \gamma P$ the corresponding discounted Markov kernel, it holds that:*

- *the stationary distribution of P_γ is the γ-discounted stationary distribution of P, i.e., $\pi_\gamma = (1 - \gamma)\nu + \gamma \pi_\gamma P = (1 - \gamma)\nu(I - \gamma P)^{-1}$;*

- *let $\Lambda\left(P^T\right) = \{1, \beta_2, \ldots, \beta_{|\mathcal{X}|}\}$ with $1 \geqslant |\beta_2| \geqslant \cdots \geqslant |\beta_{|\mathcal{X}|}|$ be the left spectrum of P, then the left spectrum of P_γ is given by:*

$$\Lambda\left(P_\gamma^T\right) = \left\{1, \gamma\beta_2, \ldots, \gamma\beta_{|\mathcal{X}|}\right\}.$$

Proof. We start with the first statement. We have to prove that π_γ is a left eigenvalue of P_γ:

$$\pi_\gamma P_\gamma = \pi_\gamma \left((1 - \gamma)\mathbf{1}\nu_0 + \gamma P\right) = (1 - \gamma) \underbrace{\pi_\gamma \mathbf{1}}_{=1} \nu + \gamma \pi_\gamma P$$

$$= (1 - \gamma)\nu \underbrace{\left(I + \gamma(I - \gamma P)^{-1}P\right)}_{=(I - \gamma P)^{-1}} = (1 - \gamma)\nu(I - \gamma P)^{-1} = \pi_\gamma.$$

We move to the second statement. For $\gamma = 1$, the statement hold. Thus, we limit to $\gamma \in [0,1)$. From Lemma 1 of Haveliwala and Kamvar (2003), we have that $|\beta_2\left(P_\gamma^T\right)| < 1$ for $\gamma \in [0,1)$. Since P_γ is a row stochastic matrix, i.e., $P_\gamma \mathbf{1} = \mathbf{1}$, we have that $\mathbf{1}$ is a right eigenvector associated to eigenvalue 1. From page 4 of Wilkinson (1971), we have that left and right eigenvectors associated to different eigenvalues are orthogonal. In particular, we take $\mathbf{1}$ as right eigenvector associated to eigenvalue 1 and x_i with $i > 1$ as left eigenvector of P_γ associated to eigenvalue β_i. As $|\beta_i| < 1$, we have that $x_i \mathbf{1} = \mathbf{0}$. Thus, we have:

$$x_i P_\gamma = (1 - \gamma)x_i \mathbf{1}\nu + \gamma x_i P = \gamma x_i P = \beta_i x_i \implies x_i P = \frac{\lambda_i}{\gamma}x_i.$$

Thus, x_i is also eigenvector of P associated to eigenvalue $\frac{\beta_i}{\gamma}$. We get the result by a change of variable. $\qquad \square$

Lemma A.5. *Let $f(x) = \gamma^x + a\sqrt{x}$ with $\gamma \in [0.3, 1]$, $a > 0$, and $x \geqslant 1$. Then, $\min_{x \geqslant 1} f(x) \geqslant \frac{a}{\sqrt{6(1-\gamma)}}$.*

Proof. We perform the following variable substitution:

$$y = x \log \frac{1}{\gamma} \quad \implies \quad f(y) = e^{-y} + \frac{a}{\sqrt{\log \frac{1}{\gamma}}}\sqrt{y} = e^{-y} + b\sqrt{y},$$

with $b = \frac{a}{\sqrt{\log \frac{1}{\gamma}}}$. We now vanish the derivative of $f(y)$:

$$\frac{\partial f}{\partial y}(y) = -e^{-y} + \frac{b}{2\sqrt{y}} = 0 \quad \implies \quad -2ye^{-2y} = -\frac{b^2}{2}$$

$$\implies \quad -2y_{0,-1} = W_{0,-1}\left(-\frac{b^2}{2}\right)$$

$$\implies \quad y_{0,-1} = -\frac{1}{2}W_{0,-1}\left(-\frac{b^2}{2}\right)$$

$$\implies \quad x_{0,-1} = -\frac{1}{2\log \frac{1}{\gamma}}W_{0,-1}\left(-\frac{a^2}{2\left(\log \frac{1}{\gamma}\right)^2}\right),$$

where $W_{0,-1}$ denote the two branches of the Lambert function. Clearly, such solutions exist provided that $-b^2/2 \geqslant -1/e$, i.e., for $a \leqslant \sqrt{2\log(1/\gamma)/e}$. If the solutions exists, we know from the Lambert function that for $-1/e \leqslant z \leqslant 0$, we have $W_0(z) \geqslant W_{-1}(z)$. Thus, $y_{-1} \geqslant y_0$. Furthermore, we have $\frac{\partial f}{\partial y}(y)|_{y\to 0} = +\infty$ and $\frac{\partial f}{\partial y}(y)|_{y\to +\infty} = +\infty$. Thus, either both stationary points are inflction points or y_0 is a local maximum and y_{-1} is a local minimum. Let us consider the second derivative $\frac{\partial^2 f}{\partial y^2}(y) = e^{-y} - \frac{a}{4y^{3/2}}$, that clearly changes sign at least once for $y \geqslant 0$. Thus, we have that y_0 is a local maximum and y_{-1} is a local minimum. For our purposes, thus, we retain y_{-1} only. To get more usable expressions, we consider the bounds on the Lambert function provided in (Chatzigeorgiou, 2013, Theorem 1), for $0 \leqslant z \leqslant 1/e$:

$$-W_{-1}(-z) \begin{cases} \leqslant -\log z + \sqrt{2(-1-\log z)} \leqslant -2\log z \\ \geqslant \frac{1}{3} - \frac{2}{3}\log z + \sqrt{2(-1-\log z)} \geqslant -\frac{2}{3}\log z \end{cases}.$$

Thus, we have:

$$f(y_{-1}) \geqslant e^{\log \frac{b^2}{2}} + b\sqrt{-\frac{1}{3}\log \frac{b^2}{2}} \geqslant \frac{b}{\sqrt{3}},$$

since $e^{\log \frac{b^2}{2}} = \frac{b^2}{2} \geqslant 0$ and, since $b^2/3 \leqslant \frac{1}{e}$ it follows that $-\log \frac{b^2}{2} \geqslant 1$. By replacing the value of b defined in terms of a and γ, we obtain:

$$f(x_{-1}) \geqslant \frac{a}{\sqrt{3\log \frac{1}{\gamma}}} \geqslant \frac{a}{\sqrt{6(1-\gamma)}},$$

for $\gamma \geqslant 0.3$.

Instead, in the case, $a < \log(1/\gamma)/\sqrt{e}$, the minimum is attained for $x = 0$, i.e., $f(0) = 1$. □