Truncating Trajectories in Monte Carlo Reinforcement Learning

Riccardo Poiani [1] **Alberto Maria Metelli** [1] **Marcello Restelli** [1]

Abstract

In Reinforcement Learning (RL), an agent acts in an unknown environment to maximize the expected cumulative discounted sum of an external reward signal, i.e., the expected return. In practice, in many tasks of interest, such as policy optimization, the agent usually spends its interaction budget by collecting episodes of *fixed length* within a simulator (i.e., Monte Carlo simulation). However, given the discounted nature of the RL objective, this data collection strategy might not be the best option. Indeed, the rewards taken in early simulation steps weigh exponentially more than future rewards. Taking a cue from this intuition, in this paper, we design an a-priori budget allocation strategy that leads to the collection of trajectories of different lengths, i.e., *truncated*. The proposed approach provably minimizes the width of the confidence intervals around the empirical estimates of the expected return of a policy. After discussing the theoretical properties of our method, we make use of our trajectory truncation mechanism to extend Policy Optimization via Importance Sampling (POIS, Metelli et al., 2018) algorithm. Finally, we conduct a numerical comparison between our algorithm and POIS: the results are consistent with our theory and show that an appropriate truncation of the trajectories can succeed in improving performance.

1. Introduction

In Reinforcement Learning (RL, Sutton & Barto, 2018), an agent acts in an unknown, or partially known, environment to maximize the expected cumulative discounted sum of an external reward signal, referred to as expected return. This abstract scenario models a large variety of sequential decision-making problems (e.g., Mnih et al., 2016; Casas, 2017; Schulman et al., 2017; Shi et al., 2020), and, consequently, is constantly gaining attention from the community. A particular appealing feature is that RL is fully data-driven. Indeed, the designer of the learning system only needs to let the agent interacts with the environment to gather experience on the task of interest, and no additional expert knowledge on the problem is required.

However, given the well-known data inefficiency of RL algorithms, in real-world scenarios, a simulator is commonly adopted, and the agent interacts with it, usually in parallel over a cluster of machines to gather knowledge (e.g., performance and gradient estimates) on the task being solved (e.g., Espeholt et al., 2018; Liang et al., 2018). Furthermore, since the goal consists on estimating/maximizing an infinite sum of rewards, in practice, the designer usually chooses a sufficiently large horizon T, so that the agent will gather information up to time T via Monte Carlo simulation (Owen, 2013), after which the state of the system is reset to a (possibly stochastic) initial state. Although there exists alternatives, such as Temporal Difference (TD, Sutton & Barto, 2018) methods, that do not require a finite horizon nor a reset possibility, a large variety of successful RL approaches still rely on Monte Carlo evaluation. Indeed, differently from TD methods, Monte Carlo approaches can be transparently applied to non-Markovian environments, as often happens in real-world domains. This is, indeed, the usual case of policy search methods (e.g., Williams, 1992; Baxter & Bartlett, 2001; Lillicrap et al., 2015; Schulman et al., 2015; 2017; Metelli et al., 2018; Cobbe et al., 2021). While these algorithms can differ across a large number of dimensions (see, for instance, Metelli et al. (2018) for an in-depth taxonomy), most of them share a common aspect: the evaluation and optimization of the objective function are performed by collecting, via Monte Carlo simulation, a batch of K episodes of length T each. In this sense, they allocate the budget of $\Lambda = KT$ transitions *uniformly* w.r.t. the horizon.

However, given the discounted nature of the RL objective, coupled with the fact that, in practice, we have to estimate the expected return with sample means, *is this uniform-in-the-horizon budget allocation strategy the best option?* Indeed, the discounted objective weighs each reward collected at step t with the factor γ^t, and, consequently, the early interaction steps weigh exponentially more than the

[1]Diparimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy. Correspondence to: Riccardo Poiani <riccardo.poiani@polimi.it>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

late ones. Building on this observation, in this work, we aim at answering the previous question by investigating alternative and non-uniform budget allocation strategies. More specifically, we tackle the problem from a worst-case scenario, which is agnostic to the underlying MDP and policy to evaluate/optimize, and we investigate whether it is possible to design an alternative schedule of trajectories' length that comes with desirable robustness properties w.r.t. to the usual uniform-in-the-horizon scheme. In other words, we aim at understanding whether the possibility of resetting trajectories, which is usually available in a large variety of RL simulators, and indispensable for Monte Carlo simulation, can successfully be exploited to increase some quality index related to the estimation accuracy.

Contributions and Outline After introducing the background (Section 2), we consider the problem of estimating the expected return of a policy via trajectory-based Monte Carlo simulation with a finite budget Λ of transitions (Section 3). For presentation purposes, we first focus on the on-policy setting, and propose a novel estimator for the expected return, which uses trajectories of *different lengths*, i.e., *truncated*. Then, to investigate alternative budget allocation strategies from our worst-case perspective, we provide a generalization of the Höeffding confidence intervals (Boucheron et al., 2003) to our estimator, and we frame our goal as finding the trajectories' length schedule that minimizes such intervals. In this sense, we design an approximately optimal fixed strategy that provably minimizes the width of these confidence intervals around the empirical mean of the RL objective. As our theory verifies, our schedule leads to collecting trajectories of different lengths. We then analyze our solution from a Probabilistic Approximately Correct (PAC, Even-Dar et al., 2002) perspective and discuss its benefits, in terms of the resulting PAC bound, w.r.t. the usual uniform-in-the-horizon approach. We conclude Section 3 by extending our approach to the more challenging off-policy evaluation setting. To this end, we minimize a generalization of the off-policy confidence intervals presented in Metelli et al. (2018). Then, while in principle, one could try to extend any algorithm that alternates steps of MC simulations with steps of optimization with the proposed schedule of trajectories, we leverage it to extend the Policy Optimization via Importance Sampling (POIS) algorithm (Metelli et al., 2018). This choice is justified by the fact that the confidence intervals optimized in Section 3 are explicitly employed in POIS to quantify the uncertainty injected in the estimation and to build a surrogate objective function of the expected return, which is then optimized via gradient methods. For this reason, in Section 4, we make use of the truncated off-policy estimator, together with our optimized confidence intervals, to extend POIS by incorporating our trajectory truncation mechanism, presenting Truncating Trajectories in Policy Optimization via Importance Sampling (TT-POIS). Finally, in Section 5, we empirically compare our algorithm and POIS across multiple control domains, varying the discount factor and the available budget. Our results are consistent with our theory and show that an appropriate truncation of the trajectories succeeds in improving performance.

2. Preliminaries

In this section, we provide the necessary backgrounds and notations that will be used throughout the rest of the article.

Markov Decision Process A discrete-time Markov Decision Process (MDP, Puterman, 1990) is defined as a tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma, \nu)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function that assigns the reward $R(s, a)$ for taking action a in state s, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$[1] is the transition kernel that specifies the probability distribution $P(\cdot|s, a)$ over the next state when taking action a in state s, $\gamma \in (0, 1)$ is the discount factor, and $\nu \in \Delta(\mathcal{S})$ is the initial-state distribution. The behavior of the agent is defined by a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that provides a mapping between states and distributions over action. We define a trajectory of length h as $\boldsymbol{\tau}_h \coloneqq (s_0, a_0, \ldots, s_{T-1}, a_{h-1}, s_h)$, i.e., a sequence of state-action pairs of length h, and we define the trajectory return as $G(\boldsymbol{\tau}_h) \coloneqq \sum_{t=0}^{h-1} \gamma^t R(s_t, a_t)$. Each trajectory of length h belongs to a trajectory space denoted with $\mathcal{T}_h$. The performance of the agent is evaluated in terms of *expected return*, i.e., the expected cumulative discounted sum of rewards over the estimation horizon T:[2] $J(\pi) \coloneqq \mathbb{E}_\pi [G(\boldsymbol{\tau}_T)]$, where the expectation is taken w.r.t. the stochasticity of the policy, the environment, and the initial-state distribution.

Policy Optimization For what concerns optimization tasks, we focus on the case in which the agent's policy belongs to a parametric differentiable policy space $\Pi_\Theta \coloneqq \{\pi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^u\}$. In this context, the expected return of any policy $\pi_{\boldsymbol{\theta}}$ is usually expressed as an integral over the trajectory space $\mathcal{T}_T$. In particular, the agent's maximization objective can be re-written as:

$$\underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} J(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\pi_{\boldsymbol{\theta}}} [G(\boldsymbol{\tau}_T)] = \int_{\mathcal{T}_T} p(\boldsymbol{\tau}_T|\boldsymbol{\theta}, T) G(\boldsymbol{\tau}_T) \mathrm{d}\boldsymbol{\tau}_T, \quad (1)$$

where $p(\boldsymbol{\tau}_h|\boldsymbol{\theta}, h) \coloneqq \nu(s_0) \prod_{t=0}^{h-1} \pi_{\boldsymbol{\theta}}(a_t|s_t) P(s_{t+1}|s_t, a_t)$ is the trajectory density function for trajectories of length

[1] We denote with $\Delta(\mathcal{X})$ the set of probability distributions over a generic set $\mathcal{X}$.

[2] As usual in the policy gradient literature (see e.g., Papini et al., 2019), we consider the infinite-horizon discounted MDP model in our setting, but a finite horizon T when introducing estimators. This is justified by the fact that, if $T = \mathcal{O}\left(\frac{1}{1-\gamma} \log \frac{1}{\epsilon}\right)$, the expected return with horizon T is ϵ-close to the infinite-horizon case (Kakade, 2003).

h. A typical approach for solving (1) is to use stochastic gradient ascent methods. For instance, the well-known REINFORCE algorithm (Williams, 1992), at each iteration, spends its interaction budget $\Lambda = KT$ in collecting K i.i.d. trajectories of length T, i.e., $\{\boldsymbol{\tau}_T^{(i)}\}_{i=1}^K$, and applies the update rule $\boldsymbol{\theta}' = \boldsymbol{\theta} + \alpha\hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$, where:

$$\hat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{K}\sum_{i=1}^{K}\left(\sum_{t=0}^{T-1}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}\left(a_t^{(i)}|s_t^{(i)}\right)\right)G(\boldsymbol{\tau}_T^{(i)})$$

represents the estimator of the *policy gradient* (Sutton et al., 1999) and $\alpha \geq 0$ is the step size.

Importance Sampling Let P and Q be two probability measures defined over a measurable space $(\mathcal{X}, \mathcal{F})$, and assume that $P \ll Q$, i.e., P is absolutely continuous w.r.t. Q. Let p and q be the density functions corresponding to P and Q respectively. In this setting, Importance Sampling (IS, Owen, 2013) is a statistical tool that allows estimating expectation $\mu = \mathbb{E}_{x\sim P}[f(x)]$ of a bounded function f (i.e., $\|f\|_\infty < +\infty$) under the target distribution P with samples collected with the behavioral distribution Q. More specifically, the IS estimator corrects the distribution mismatch via the *importance weights* $\omega_{P/Q}(x) = p(x)/q(x)$:

$$\hat{\mu}_{P/Q} = \frac{1}{K}\sum_{i=1}^{K}\omega_{P/Q}(x_i)f(x_i), \tag{2}$$

where $\{x_i\}_{i=1}^K \sim Q$. The moments of the importance weights can be expressed in terms of the exponentiated Rényi divergence. More specifically, let $\alpha \in [0, +\infty]$, the α-Rényi divergence between P and Q is defined as:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\log\int_{\mathcal{X}} q(x)\left(\frac{p(x)}{q(x)}\right)^\alpha \mathrm{d}x.$$

We define $d_\alpha(P\|Q) \coloneqq \exp\left(D_\alpha(P\|Q)\right)$ as the exponentied α-Rényi divergece, then $\mathbb{E}_{x\sim Q}\left[\omega_{P/Q}(x)^\alpha\right] = d_\alpha(P\|Q)^{\alpha-1}$. The second order moments can be used to construct the following confidence intervals on the target estimation (Metelli et al., 2018) that holds with probability at least $1-\delta$:

$$\mathbb{E}_{x\sim Q}[f(x)] \geq \hat{\mu}_{P/Q} - \|f\|_\infty\sqrt{\frac{(1-\delta)d_2(P\|Q)}{\delta K}} \tag{3}$$

3. Truncating Trajectoris in Monte Carlo Evaluation

In this section, we provide the theoretical groundings behind truncating trajectories in Monte Carlo RL, with a specific focus on the problem of estimating the discounted return. Before diving into the details of our approach, we first formally specify how an agent makes use of its interaction budget. For this purpose, we introduce the novel concept of *Data Collection Strategy* (DCS).

Definition 3.1 (Data Collection Strategy)**.** A *Data Collection Strategy* (DCS) for a transition budget $\Lambda \in \mathbb{N}$ is defined as a T-dimensional vector $\boldsymbol{m} \coloneqq (m_1, \ldots, m_T)$ such that $m_h \in \mathbb{N}$ for all $h \in \{1, \ldots, T\}$, and $\sum_{h=1}^T m_h h = \Lambda$.

More specifically, m_h represents the number of trajectories of length h that the agent collects in the environment. We notice that there is a tight relationship between $\boldsymbol{m}$ and the total number of samples that the agent collects at step t. In particular, let $\boldsymbol{n} \coloneqq (n_0, \ldots, n_{T-1})$ be the T-dimensional vector, where each component n_t represents the number of samples collected at time t; then, we have that $n_t = m_t$, if $t = T-1$, and $n_t = n_{t+1} + m_{t+1}$ otherwise.[3] It follows that, given $\boldsymbol{m}$, $\boldsymbol{n}$ is uniquely identified, and vice versa; for this reason, in the rest of this paper, we will use the most convenient symbol depending on the context. Finally, we remark that each DCS corresponds to $p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})$, which represents the density function of the data generation process of the trajectories collected under policy $\pi_{\boldsymbol{\theta}}$ following $\boldsymbol{m}$, namely $\mathcal{D} \coloneqq \{\{\boldsymbol{\tau}_h^{(i)}\}_{i=1}^{m_h}\}_{h=1}^T$.

3.1. On-Policy Data Collection Strategy

Let us now consider the on-policy problem of estimating $J(\boldsymbol{\theta})$ with trajectories collected via Monte Carlo simulation using $\pi_{\boldsymbol{\theta}}$. We begin by investigating whether, having fixed an arbitrary DCS, it is possible to build unbiased estimators for $J(\boldsymbol{\theta})$.[4] The answer turns out to be positive for a restricted class of DCSs, namely the ones for which $m_T \geq 1$ holds. Intuitively, this condition ensures that the agent gathers at least one sample for each interaction step $t \in \{0, \ldots, T-1\}$. Thus, for any DCS $\boldsymbol{m}$, we design the following estimator:

$$\hat{J}_{\boldsymbol{m}}(\boldsymbol{\theta}) = \sum_{h=1}^{T}\sum_{i=1}^{m_h}\sum_{t=0}^{h-1}\gamma^t\frac{R(s_t^{(i)}, a_t^{(i)})}{n_t}, \tag{4}$$

To provide an interpretation, we notice that, given $\mathcal{D} \sim p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})$, Equation (4) sums over the collected trajectories of different lengths (i.e., the two external summations) a rescaled empirical truncated return in which each reward at step t is *properly* divided by n_t, i.e., the number of samples gathered at step t. Intuitively, this rescaling is required to prevent the estimate to be biased toward the steps for which n_t is larger. Furthermore, this estimator has several interesting properties. First of all, as already anticipated, as long as $m_T \geq 1$, it provides an unbiased estimate of $J(\boldsymbol{\theta})$, namely $\mathbb{E}_{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}\left[\hat{J}_{\boldsymbol{m}}(\boldsymbol{\theta})\right] = J(\boldsymbol{\theta})$ (proof in Appendix B). Moreover, given the uniform-in-the-horizon

[3] Notice that this implies that $n_t \geq n_{t+1}$.

[4] Notice that a naïve Monte Carlo estimator such as $\frac{1}{\Lambda}\sum_{h=1}^T\sum_{i=1}^{m_h}\sum_{t=0}^{h-1}\gamma^t R(s_t^{(i)}, a_t^{(i)})$ is, in general, biased.

DCS, i.e., $\boldsymbol{m} = \left(0, \dots, 0, \frac{\Lambda}{T}\right)$, Equation (4) recovers the usual Monte Carlo on-policy estimator of $J(\boldsymbol{\theta})$. As we shall later see, these properties will also naturally extends to the off-policy estimation problem.

At this point, our main objective can be framed as finding the best possible DCS among the ones that preserve the unbiasedness property. In this sense, we need to define a proper index to evaluate the candidates. In this work, we take a worst-case scenario w.r.t. the underlying MDP and policy, and we choose to minimize confidence intervals around the estimated expected return. More specifically, we derive the following generalization of the Höeffding confidence intervals (Boucheron et al., 2003) that holds for a generic DCS (proof in Appendix B). [5]

Proposition 3.2. *Consider an optimization budget $\Lambda \geq T$, a generic DCS $\boldsymbol{m}$ such that $m_T \geq 1$ and $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$ it holds that:*

$$\left|\hat{J}_{\boldsymbol{m}}(\boldsymbol{\theta}) - J(\boldsymbol{\theta})\right| \leq \sqrt{\frac{1}{2}\log\left(\frac{2}{\delta}\right)\sum_{t=0}^{T-1}\frac{c_t}{n_t}}, \tag{5}$$

where $c_t = \frac{\gamma^t(\gamma^t + \gamma^{t+1} - 2\gamma^T)}{1-\gamma}$.

As we can notice, Equation (5) is always well-defined. Indeed, $m_T \geq 1$ implies $n_t \geq 1$ for all $t \in \{0, \dots, T-1\}$. Furthermore, a single term within the summation $\frac{c_t}{n_t}$ relates the width of the confidence intervals w.r.t. the number of samples gathered at timestep t. More specifically, we notice that c_t is a decreasing function of time. Intuitively, if we are given a fixed budget Λ, we expect that, to minimize Equation (5), more samples should be allocated at the beginning of the horizon, corroborating our initial intuition, i.e., the convenience of truncating trajectories. Moreover, we notice that the discount factor γ plays a crucial role in the expression of c_t. The lower γ, the faster the aforementioned decreasing rate, meaning that, for small γs, a larger portion of the budget Λ will be allocated to earlier interaction steps when minimizing Equation (5). Finally, it is possible to verify that, when Proposition 3.2 is applied with the uniform DCS, we recover the usual Höeffding confidence intervals for the Monte Carlo estimation of $J(\boldsymbol{\theta})$. Therefore, given a fixed budget Λ, if we are able to find the DCS that minimizes Equation (5), we implicitly obtain a *robustness* property w.r.t. the uniform-in-the-horizon strategy. In order to find the DCS that minimizes Equation (5), we formulate the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{n}} \quad & f(\boldsymbol{n}) \coloneqq \sqrt{\frac{1}{2}\log\left(\frac{2}{\delta}\right)\sum_{t=0}^{T-1}\frac{c_t}{n_t}} \\ \text{s.t.} \quad & \sum_{t=0}^{T-1} n_t = \Lambda \\ & n_t \geq n_{t+1}, \quad \forall t \in \{0, \dots, T-2\} \\ & n_t \in \mathbb{N}_+, \quad \forall t \in \{0, \dots, T-1\} \end{aligned} \tag{6}$$

where the constraint $n_t \geq n_{t+1}$ arises from the aforementioned relationships between $\boldsymbol{m}$ and $\boldsymbol{n}$. Problem (6) is a non-linear integer program that, in principle, could be addressed by means of complex solvers. However, such an approach would fail to provide an interpretable result (e.g., closed-form expression) and, thus, would be of little interest for statistical analysis. For this reason, we follow a different path and derive an analytical expression for an approximately optimal DCS, whose form arises from solving a convex relaxation of (6) in which the integer constraint on n_t is dropped and, then, the obtained optimal relaxed solution is rounded down and the remaining budget is allocated uniformly. For the sake of presentation, the following Theorem (proof in Appendix B) summarizes our result for a sufficiently large budget $\Lambda \geq \Lambda_0$. We refer the reader to Appendix B for the exact expression of Λ_0 and for the symmetric version of Theorem 3.3 that holds when $T \leq \Lambda < \Lambda_0$.

Theorem 3.3. *Consider an optimization budget $\Lambda \geq \Lambda_0$, let $\boldsymbol{n}^*$ be the optimal solution of* (6)*. Let $g_t = \frac{\sqrt{c_t}}{\sum_{i=0}^{T-1}\sqrt{c_i}}\Lambda$, and let $k = \Lambda - \sum_{t=0}^{T-1}\lfloor g_t \rfloor$. Define the t-th component of the approximately optimal DCS $\tilde{\boldsymbol{n}}^*$ as $\tilde{n}_t^* \coloneqq \lfloor g_t \rfloor + \mathbf{1}\{t < k\}$. Then, it holds that:*

$$f(\boldsymbol{n}^*) \leq f(\tilde{\boldsymbol{n}}^*) \leq \sqrt{2} f(\boldsymbol{n}^*). \tag{7}$$

Theorem 3.3 deserves some comments. First of all, it provides a closed form expression for an approximately optimal DCS $\tilde{\boldsymbol{n}}^*$. Indeed, from Equation (7) we can infer that, up to constant factors, $\tilde{\boldsymbol{n}}^*$ achieves the same confidence intervals as the true optimal DCS $\boldsymbol{n}^*$ that minimizes (6) (i.e., $f(\tilde{\boldsymbol{n}}^*) = \Theta(f(\boldsymbol{n}^*))$). Now, let us focus on the expression of a single term $\tilde{n}_t^*$, whose shape in visualized in Figure 1. [6] Neglecting the indicator function and the floor, which both arise from technicalities in the analysis, the most relevant term is given by g_t. As one can notice, g_t partitions the available budget Λ among the different timesteps with a proportion of $\sqrt{c_t}$, which is an exponentially decreasing function of time and whose decrease rate is given by γ. In this sense, the approximately optimal DCS provably truncates trajectories by allocating more samples to the initial

[5] We remark that Proposition 3.2 does not directly follow from a naïve application of the Hoeffding inequality, and some tecnical manipulations are required to obtain Equation (5).

[6] Further visualizations are provided in Appendix C.

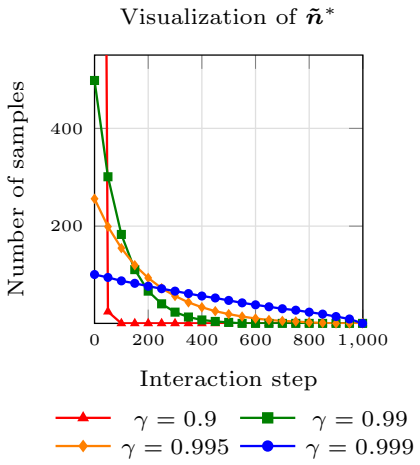

Figure 1. Visualization of $\tilde{\boldsymbol{n}}^*$ for $\Lambda = 50000$, $T = 1000$, and different values of γ. More specifically, the x-axis denotes the interaction timestep t and the y-axis reports, for each value of t, the number of steps prescribed by $\tilde{\boldsymbol{n}}^*$, namely $\tilde{n}_t^*$. As we can see, the behavior of $\tilde{n}_t^*$ is monotonically decreasing in t; furthermore, the smaller γ is, the faster the decrease rate of $\tilde{n}_t^*$ is.

steps of interactions. Moreover, the smaller γ is, the larger the amount of samples that will be allocated to earlier steps t, as aggressive discounting makes the future less relevant.

To conclude this section, we remark that Theorem 3.3 provably shows that truncating trajectories can successfully minimize the confidence intervals around the estimated return, for any possible pair of MDP and target policy. However, the complexity of the expression of the approximately optimal DCS $\tilde{\boldsymbol{n}}^*$ does not allow to easily quantify the improvement w.r.t. the uniform strategy. For this reason, we resort to PAC analysis (Even-Dar et al., 2002). More specifically, given some desired confidence level $\delta \in (0, 1)$ and accuracy $\epsilon > 0$, we aim at answering the following question: which is the minimum amount of budget Λ such that $|\hat{J}_{\boldsymbol{m}}(\boldsymbol{\theta}) - J(\boldsymbol{\theta})| \leq \epsilon$ holds with probability at least $1 - \delta$? It is easy to see that, for the uniform DCS, $\Lambda = \mathcal{O}\left(\frac{T \log(2/\delta)}{(1-\gamma)^2 \epsilon^2}\right)$ is sufficient for enforcing $|\hat{J}_{\boldsymbol{m}}(\boldsymbol{\theta}) - J(\boldsymbol{\theta})| \leq \epsilon$. For our approximately optimal DCS, instead, we derive the following result:

Theorem 3.4. *Let $\delta \in (0, 1)$ and $\epsilon > 0$ such that $8T\epsilon^2 \leq \log(2/\delta)c_0$ holds. Then, with probability at least $1 - \delta$, $|\hat{J}_{\tilde{\boldsymbol{m}}^*}(\boldsymbol{\theta}) - J(\boldsymbol{\theta})| \leq \epsilon$ holds provided that:*

$$\Lambda = \mathcal{O}\left(\min\left\{\frac{T \log(2/\delta)}{(1-\gamma)^2 \epsilon^2}, \frac{\log(2/\delta)}{(1-\gamma)^3 \epsilon^2}\right\}\right) \tag{8}$$

Theorem 3.4 reveals a PAC bound under an assumption on the relationship between ϵ, δ, and T; technically, this is only needed to guarantee that $\Lambda \geq 2T$ holds, which is clearly a mild condition. That being said, Theorem 3.4, first of all, shows a robustness property of $\tilde{\boldsymbol{m}}^*$ w.r.t. the uniform DCS. Moreover, it improves the standard result whenever $T > \mathcal{O}\left((1-\gamma)^{-1}\right)$, as shown in the following example.

Example 3.5. Suppose that the agent is interested in estimating the infinite-horizon expected discounted return using a finite horizon T that guarantees that the final estimate has bias bounded by $1/\exp\left((1-\gamma)^{-1}\right)$. In this case, we need to select $T = \mathcal{O}\left((1-\gamma)^{-2}\right)$ (Kakade, 2003), and the improvement of our approximately optimal DCS is given by a factor $\mathcal{O}\left((1-\gamma)^{-1}\right)$ factor. This result should not be surprising. Indeed, intuitively, if the horizon increases, the difference between the optimal DCS and the uniform one increases as well (see the expression of $\tilde{n}_t^*$ vs Λ/T).

At this point, we are ready to extend our result to the off-policy estimation problem.

3.2. Off-Policy Data Collection Strategy

Consider the off-policy problem of estimating $J(\bar{\boldsymbol{\theta}})$ with trajectories collected via Monte Carlo simulation using a possibly different policy $\pi_{\boldsymbol{\theta}}$. For an arbitrary DCS $\boldsymbol{m}$, we extend Equation (4) by proposing the following estimator:

$$\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) = \sum_{h=1}^{T} \sum_{i}^{m_h} \omega_{\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}}(\boldsymbol{\tau}_h^{(i)}) \sum_{t=0}^{h-1} \gamma^t \frac{R(s_t^{(i)}, a_t^{(i)})}{n_t}, \tag{9}$$

where $\omega_{\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}}(\boldsymbol{\tau}_h) = \prod_{t=0}^{h-1} \frac{\pi_{\bar{\boldsymbol{\theta}}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}}(a_t|s_t)}$ is the importance weight for a trajectory of length h. Equation (9) enjoys the same properties discussed for Equation (4); the only difference, indeed, stands in $\omega_{\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}}(\boldsymbol{\tau}_h)$, whose purpose is taking into account the distribution shift. Following the same rationale of the on-policy setting, we derive a generalization for the off-policy confidence intervals of Metelli et al. (2018) for the discounted off-policy return $J(\bar{\boldsymbol{\theta}})$ that holds for a generic DCS $\boldsymbol{m}$.[7] More specifically, we prove the following result (proof in Appendix B).[8]

Theorem 3.6. *Consider $\pi_{\bar{\boldsymbol{\theta}}}, \pi_{\boldsymbol{\theta}} \in \Pi_\Theta$ such that $\pi_{\bar{\boldsymbol{\theta}}}(\cdot|s) \ll \pi_{\boldsymbol{\theta}}(\cdot|s)$ a.s. for all $s \in \mathcal{S}$. Consider an optimization budget $\Lambda \geq T$ and a generic DCS $\boldsymbol{m}$. Then, with probability at least $1 - \delta$ it holds that:*

$$J(\bar{\boldsymbol{\theta}}) \geq \hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) - \sqrt{\beta_\delta \sum_{h=1}^{T} m_h \phi_h^2 d_2(p(\cdot|\bar{\boldsymbol{\theta}}, h) \| p(\cdot|\boldsymbol{\theta}, h))}, \tag{10}$$

where $\beta_\delta = \frac{1-\delta}{\delta}$ and $\phi_h \coloneqq \sum_{t=0}^{h-1} \frac{\gamma^t}{n_t}$.

[7] We remark that for Equation (9), that uses IS, we cannot easily apply the Höeffding's inequality since the importance weight distribution might be heavy tailed (Lugosi & Mendelson, 2019).

[8] Theorem 3.6 makes use of Cantelli's inequality and provides one-sided tail bounds. Two-sided tail bounds can be straightforwardly derived by using Chebyshev's inequality.

However, Equation (10) is of little practical use to derive a DCS. Indeed, minimizing Equation (10) as a function of $\boldsymbol{m}$ entails computing the Rényi divergence over the trajectory space. This, in turn, requires both to compute the approximation of a complex integral, and, for stochastic environments, the knowledge of the transition kernel P of the underlying MDP (Metelli et al., 2018). Therefore, to derive a tractable expression, we further bound each term as $d_2(p(\cdot|\bar{\boldsymbol{\theta}}, h)\|p(\cdot|\boldsymbol{\theta}, h)) \leq d_2(p(\cdot|\bar{\boldsymbol{\theta}}, T)\|p(\cdot|\boldsymbol{\theta}, T))$, thus leading to:

$$\sqrt{\beta_\delta d_2(p(\cdot|\bar{\boldsymbol{\theta}}, T)\|p(\cdot|\boldsymbol{\theta}, T)) \sum_{t=0}^{T-1} \frac{c_t}{n_t}}. \tag{11}$$

However, it is easy to verify that, finding the DCS that minimizes this new expression, leads to an optimization problem with the same structure of Problem (6). Indeed, $2\log(2/\delta)$ and $\beta_\delta d_2(p(\cdot|\bar{\boldsymbol{\theta}}, T)\|p(\cdot|\boldsymbol{\theta}, T))$ can be seen as constants that do not impact the result of the optimization. For this reason, it is possible to derive an equivalent of Theorem 3.3 for the off-policy setting that we defer to Appendix B. Notice, however, that the form of the approximately optimal DCS $\tilde{\boldsymbol{m}}^*$ is left unchanged, and, consequently, all the previous comments about the on-policy solution extends to the off-policy setting as well. Consequently, one can obtain also an equivalent of Theorem 3.4 that expresses PAC bounds for the off-policy setting. Further details on this point are provided in Appendix B.

To conclude, we remark that, although a further upper bound has been applied to obtain Equation (11), once the data has been collected, one can still make use of the tighter bound of Equation (10) to obtain a confidence interval on $J(\bar{\boldsymbol{\theta}})$. Indeed, since $d_2(p(\cdot|\bar{\boldsymbol{\theta}}, h)\|p(\cdot|\boldsymbol{\theta}, h)) \leq d_2(p(\cdot|\bar{\boldsymbol{\theta}}, T)\|p(\cdot|\boldsymbol{\theta}, T))$, this implies a further source of improvement w.r.t. the uniform strategy.

3.3. Discussion

We now discuss the choice of our confidence interval metrics to optimize the DCS. As we have seen, our method provably minimizes confidence intervals on the expected discounted return of a given policy. More specifically, the choice of the confidence intervals that we adopt, together with the methodology that we present, leads to a novel *fixed* DCS (i.e., $\tilde{\boldsymbol{m}}^*$) that can be adopted for estimation purposes. Minimizing confidence intervals is well-known to be a robust solution against heavy-tailed distributions (Lugosi & Mendelson, 2019), and, consequently, our work comes with desired statistical properties that hold for *any* possible pair of MDP and target policy. Moreover, given that the proposed DCS is pre-determined, it nicely fits situations where the agent collects its experience (i.e., spends Λ) in parallel over a cluster of machines, which is a typical scenario for policy gradient methods. At this point, one might object that there might exist MDPs in which, intuitively, truncating trajectories is a sub-optimal solution. In particular, suppose that the agent gathers rewards different from 0 only in the last interaction step (e.g., a *goal-based* problem). In this situation, we can imagine that the uniform strategy should be preferred over any other allocation strategy, even in a discounted setting.[9] Our approach does not capture this *problem-dependent* feature since it is designed to be agnostic w.r.t. the underlying structure of the MDP and target policy. However, we remark that, without any sort of prior knowledge, our method provably minimizes the *worst-case* scenario. Furthermore, we also notice that when dealing with sparse rewards, γ is usually selected to be close to 1 to avoid nullifying the positive reward gathered at the end of the trajectory. In such a scenario, $\tilde{\boldsymbol{m}}^*$ tends to the uniform strategy.

4. Truncating Trajectories in Policy Optimization via Importance Sampling

In this section, we discuss how to use our approximately optimal DCS $\tilde{\boldsymbol{m}}^*$ in a policy optimization algorithm. In particular, given the result from Section 3, Policy Optimization via Importance Sampling (POIS, Metelli et al., 2018), as we shall see in a moment, turns out to be an natural choice. POIS is a recent off-policy optimization algorithm that alternates *online* interactions with the environment (i.e., data collection) with *offline* optimization. In particular, POIS first makes use of the uniform DCS to collect a batch of K episodes of length T under the current policy $\pi_{\boldsymbol{\theta}}$. Then, it searches, by gradient steps, for the next policy $\pi_{\bar{\boldsymbol{\theta}}}$ that maximizes an empirical version of the statistical surrogate for the off-policy return derived from Equation (10). Namely, the agent optimizes for:

$$\hat{J}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) - \sqrt{\beta_\delta \hat{d}_2(p(\cdot|\bar{\boldsymbol{\theta}}, T)|p(\cdot|\boldsymbol{\theta}, T)) \left(\sum_{t=0}^{T-1} \gamma^t\right)^2 \frac{T}{\Lambda}},$$

where $\hat{d}_2(p(\cdot|\bar{\boldsymbol{\theta}}, T)|p(\cdot|\boldsymbol{\theta}, T))$ is a sampled-based estimation of the Rényi divergence $d_2(p(\cdot|\bar{\boldsymbol{\theta}}, T)|p(\cdot|\boldsymbol{\theta}, T))$ (see Equation 41 in Metelli et al. (2018)). In other words, POIS limits the update step via an adaptive trust region defined by the confidence intervals on the estimation of $J(\bar{\boldsymbol{\theta}})$ given that data have been collected using a different policy $\pi_{\boldsymbol{\theta}}$.

In this work, we build on POIS, and we propose to employ our optimized DCS $\tilde{\boldsymbol{m}}^*$, together with the corresponding estimator, to build a tighter surrogate of the off-policy return. More specifically, from Equation (10), we define $\mathcal{L}_\delta(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})$

[9] We propose a variance analysis for these scenarios in Appendix B.

as our empirical objective function:

$$\mathcal{L}_\delta(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) \coloneqq \hat{J}_{\tilde{\boldsymbol{m}}^*}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) - \sqrt{\beta_\delta \sum_{h=1}^{T} \tilde{m}_h^*(\tilde{\phi}_h^*)^2 \hat{d}_2(p(\cdot|\bar{\boldsymbol{\theta}},h)|p(\cdot|\boldsymbol{\theta},h))}, \tag{12}$$

where $\tilde{\phi}_h^* = \sum_{t=0}^{h-1} \frac{\gamma^t}{\tilde{n}_t^*}$ and $\hat{d}_2(p(\cdot|\bar{\boldsymbol{\theta}},h)|p(\cdot|\boldsymbol{\theta},h))$ is a sampled-based estimation for $d_2(p(\cdot|\bar{\boldsymbol{\theta}},h)|p(\cdot|\boldsymbol{\theta},h))$. Notice that TT-POIS, in Equation (12), makes use of the *tighter* bound of Equation (10) with the approximately optimal DCS derived while optimizing Equation (11). This choice is justified by the discussion at the end of Section 3.2.

In the following, we will refer to the algorithm using Equation (12) as objective function as Truncating Trajectories in Policy Optimization via Importance Sampling (TT-POIS). The pseudo-code, together with other practical implementation details can be found in Appendix D.

We conclude by highlighting that TT-POIS can make better use of the collected data from a statistical perspective (i.e., smaller confidence intervals), suggesting that the surrogate loss will be closer to the true return $J(\bar{\boldsymbol{\theta}})$. This implies that the adaptive trust region over the parameter space defined by Equation (12) will allow for larger update steps.

5. Experiments

In this section, we numerically validate our approach by targeting the comparison between POIS and TT-POIS across multiple domains and varying both the discount factor and the available budget. In all experiments, we first tuned the hyper-parameters with POIS, and then, we applied TT-POIS to the best hyper-parameter configuration of POIS. We now present our experimental domains, followed by a discussion of the results. Further details and additional experiments are deferred to Appendix E.

Dam Control In our first experimental domain, we consider a water resource management scenario (Castelletti et al., 2010; Parisi et al., 2014; Tirinzoni et al., 2018; Liotet et al., 2022). The goal of the agent is to learn a water release policy that trades off between some external demand D (e.g., the needs of a town) and keeping the water level below a flooding threshold F. The dam is subject to an external and stochastic net inflow, that, each day, determines the amount of additional water i_t that will be stored (e.g., rain). More specifically, this inflow profile has a periodic shape defined over a period of one year; the demand, instead, is kept constant. The state of the system s_t evolves according to a simple mass balance principle, namely $s_{t+1} = \max\{s_t - a_t + i_t, 0\}$, where a_t is the amount of water that the agent intends to release at day t. The reward $R(s_t, a_t)$ is a convex combination of the two aforementioned objectives: $-c_1 \max\{0, s_t - F\}$ (i.e., flooding control) and $-c_2 \max\{0, D - a_t\}^2$ (i.e., meeting the demand), where $c_1, c_2 > 0$ are domain-dependent constants.

Reacher In the second experiment, we consider the standard continuous control problem of a two-jointed robot arm (Todorov et al., 2012), whose goal is to move the robot's end effector close to a target spawned in a random position. The reward is a combination of a control cost together with a penalization for the end effector being far from the goal.

Multi-Echelon Supply Chain Finally, we consider the problem of managing the complex inventory of a 4 stage supply chain (Hubbs et al., 2020). During each day, the agent needs to decide, for each stage, how many products it should order from its supplier (i.e., the previous stage). Once the goods have been ordered at stage i, it takes a given amount of days t_i (i.e., lead time of stage i) so that they are shipped and delivered to stage $i + 1$. Goods at the last stage are sold according to some stochastic demand. If the retailer fails to meet the demand, lost orders are backlogged, meaning they are fulfilled later but for lower profit. The agent's key challenge is trading off the uncertainty of the demand (i.e., profit) with the costs incurred for storing products in the inventory. A mathematical description of the problem can be found in Hubbs et al. (2020).

Results Figure 2 reports the average discounted return on the considered domains (mean and 95% confidence intervals of 5 runs) varying the discount factor γ. More specifically, the first row has been obtained with $\gamma = 0.999$, while the second one with $\gamma = 0.95$ (experiments with additional values of γ can be found in Appendix E.4). The considered budget per iteration is $\Lambda = 8640$ for the Dam environment, $\Lambda = 3900$ for the Supply Chain, and $\Lambda = 8000$ for the Reacher. As we can notice, independently of the value of γ, TT-POIS always performs better w.r.t. its original version POIS. It is worth noting what happens to the training curves when we change the value of γ. In these scenarios, as previously discussed, the dissimilarity between our non-uniform DCS $\tilde{\boldsymbol{m}}^*$ and the uniform one increases. Indeed, $\tilde{\boldsymbol{m}}^*$ will allocate a larger portion of the budget Λ to the initial interaction steps. This, in turn, implies a larger difference in the confidence intervals of the surrogate objective function. Consequently, as soon as we decrease γ, we observe a larger performance improvement in each of the domains. This is consistent with our theory. Due to the tighter confidence bounds, the agent is able to make better use of the collected data and takes larger update steps.

For what concerns experiments in which the budget Λ changes, we report the results in Appendix E.4 for space constraints. However, we highlight that the behavior of the algorithms does not display significant differences. TT-

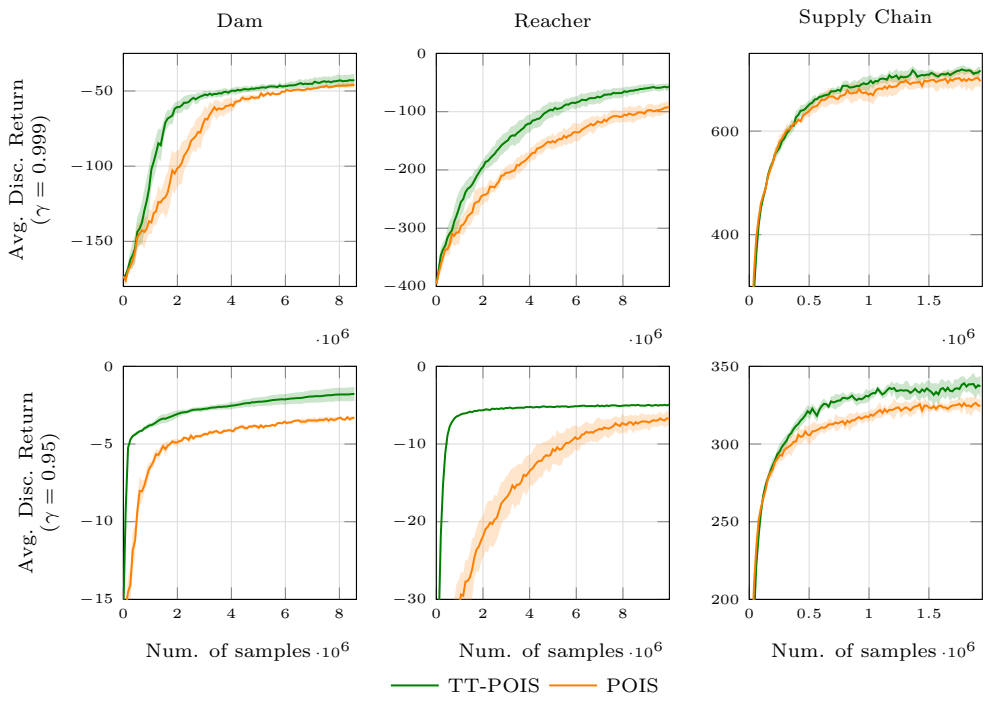

Figure 2. Experimental results (mean and 95% confidence intervals of 5 runs). The first row (resp. second row) reports average returns with $\gamma = 0.999$ (resp. $\gamma = 0.95$).

POIS always performs better than POIS, and what has been previously highlighted for Figure 2 replicates consistently.

6. Conclusions and Future Works

In this work, we focused on how to allocate the interaction budget Λ in Monte Carlo Reinforcement Learning. We started by building on the intuition that the common uniform-in-the-horizon strategy might not be the best option when discounted rewards are considered. To study the problem from a theoretical perspective, we introduced the novel concept of Data Collection Strategy (DCS), and we investigated alternative non-uniform solutions from the worst-case robust viewpoint of confidence intervals. More specifically, starting from the on-policy evaluation problem, we showed that, to minimize confidence intervals around the estimated expected return of a policy, non-uniform DCSs, which we provide in closed-form, represent a more appropriate solution, thus confirming our initial intuition. After theoretically analyzing the benefit of the proposed DCS from a PAC perspective, we further extended our reasoning to off-policy evaluation problems by generalizing the confidence bounds of Metelli et al. (2018), that are directly employed in the surrogate loss function that a recent algorithm, i.e., POIS (Metelli et al., 2018), optimizes for. We then proposed an extension of POIS, TT-POIS, that makes use of our optimized budget allocation, and we verified that it leads to performance improvements among multiple domains, and for different values of γ and Λ.

We conclude by remarking that our work roots down to a main component of RL algorithms; i.e., the interaction with the environment. More specifically, we showed that it is possible to find principled strategies that optimize the *collection* of the experience with the domain at hand. Our work, in this sense, does not close the problem but takes a first step toward this direction, thus paving the way for several exciting future works.

For example, while in this work we took a robust approach and derived a fixed data collection strategy that minimizes confidence intervals around the target return, other choices are also possible. In this sense, a complementary direction w.r.t. to the one we followed would be to find a dynamic DCS that performs *online* minimization of the MSE of some estimator while interacting with the environment. This could be possible by integrating our approach and analysis with confidence intervals that rely on empirical quantities (e.g., Maurer & Pontil, 2009), and, consequently, by designing strategies that aim at minimizing the MSE in an online fashion. Moreover, more effective strategies could be derived when restricting to specific subclasses of problems (e.g., goal-based) and leveraging their problem-dependent features.

Finally, we notice that our approach is based on Monte

Carlo data collection and, therefore, does not deeply exploit the Markovian properties of the underlying MDP. Empowering the proposed methods with TD approaches that take into consideration value functions available in actor-critic algorithms (e.g., Schulman et al., 2017) may lead to further improvements in the performance of policy-search algorithms.

Acknowledgements

This paper is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

References

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Bhatia, A., Thomas, P. S., and Zilberstein, S. Adaptive rollout length for model-based rl using model-free deep rl. *arXiv preprint arXiv:2206.02380*, 2022.

Boucheron, S., Lugosi, G., and Bousquet, O. Concentration inequalities. In *Summer school on machine learning*, pp. 208–240. Springer, 2003.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Casas, N. Deep deterministic policy gradient for urban traffic light control. *arXiv preprint arXiv:1703.09035*, 2017.

Castelletti, A., Galelli, S., Restelli, M., and Soncini-Sessa, R. Tree-based reinforcement learning for optimal water reservoir operation. *Water Resources Research*, 46(9), 2010.

Cobbe, K. W., Hilton, J., Klimov, O., and Schulman, J. Phasic policy gradient. In *International Conference on Machine Learning*, pp. 2020–2027. PMLR, 2021.

Dai, B., Nachum, O., Chow, Y., Li, L., Szepesvári, C., and Schuurmans, D. Coindice: Off-policy confidence interval estimation. *Advances in neural information processing systems*, 33:9398–9411, 2020.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, pp. 1407–1416. PMLR, 2018.

Even-Dar, E., Mannor, S., and Mansour, Y. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pp. 255–270. Springer, 2002.

Farahmand, A.-m., Nikovski, D., Igarashi, Y., and Konaka, H. Truncated approximate dynamic programming with task-dependent terminal value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

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

Hesterberg, T. C. *Advances in importance sampling*. Stanford University, 1988.

Hubbs, C. D., Perez, H. D., Sarwar, O., Sahinidis, N. V., Grossmann, I. E., and Wassick, J. M. Or-gym: A reinforcement learning library for operations research problems. *arXiv preprint arXiv:2008.06319*, 2020.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Kakade, S. M. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Kalos, M. H. and Whitlock, P. A. *Monte carlo methods*. John Wiley & Sons, 2009.

Kandasamy, K., Dasarathy, G., Poczos, B., and Schneider, J. The multi-fidelity multi-armed bandit. *Advances in neural information processing systems*, 29, 2016.

Kandasamy, K., Dasarathy, G., Oliva, J., Schneider, J., and Poczos, B. Multi-fidelity gaussian process bandit optimisation. *Journal of Artificial Intelligence Research*, 66: 151–196, 2019.

Liang, E., Liaw, R., Nishihara, R., Moritz, P., Fox, R., Goldberg, K., Gonzalez, J., Jordan, M., and Stoica, I. Rllib: Abstractions for distributed reinforcement learning. In *International Conference on Machine Learning*, pp. 3053–3062. PMLR, 2018.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Liotet, P., Vidaich, F., Metelli, A. M., and Restelli, M. Lifelong hyper-policy optimization with multiple importance sampling regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7525–7533, 2022.

Lugosi, G. and Mendelson, S. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Metelli, A. M., Papini, M., Faccio, F., and Restelli, M. Policy optimization via importance sampling. *Advances in Neural Information Processing Systems*, 31, 2018.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Mukherjee, S., Hanna, J. P., and Nowak, R. D. Revar: Strengthening policy evaluation via reduced variance sampling. In Cussens, J. and Zhang, K. (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1413–1422. PMLR, 01–05 Aug 2022.

Nguyen, N. M., Singh, A., and Tran, K. Improving model-based rl with adaptive rollout using uncertainty estimation. 2018.

Owen, A. B. Monte carlo theory, methods and examples. 2013.

Papini, M., Pirotta, M., and Restelli, M. Adaptive batch size for safe policy gradients. *Advances in Neural Information Processing Systems*, 30, 2017.

Papini, M., Pirotta, M., and Restelli, M. Smoothing policies and safe policy gradients. *arXiv preprint arXiv:1905.03231*, 2019.

Papini, M., Battistello, A., and Restelli, M. Balancing learning speed and stability in policy gradient via adaptive exploration. In *International conference on artificial intelligence and statistics*, pp. 1188–1199. PMLR, 2020.

Parisi, S., Pirotta, M., Smacchia, N., Bascetta, L., and Restelli, M. Policy gradient approaches for multi-objective sequential decision making. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 2323–2330. IEEE, 2014.

Pirotta, M., Restelli, M., and Bascetta, L. Adaptive step-size for policy gradient methods. *Advances in Neural Information Processing Systems*, 26, 2013.

Poiani, R., Metelli, A. M., and Restelli, M. Multi-fidelity best-arm identification. In *Advances in Neural Information Processing Systems*, 2022.

Puterman, M. L. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shi, W., Wei, X., Zhang, J., Ni, X., Jiang, A., Bian, J., and Liu, T.-Y. Cooperative policy learning with pre-trained heterogeneous observation representations. *arXiv preprint arXiv:2012.13099*, 2020.

Sun, W., Bagnell, J. A., and Boots, B. Truncated horizon policy search: Combining reinforcement learning & imitation learning. *arXiv preprint arXiv:1805.11240*, 2018.

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

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Tirinzoni, A., Sessa, A., Pirotta, M., and Restelli, M. Importance weighted transfer of samples in reinforcement learning. In *International Conference on Machine Learning*, pp. 4936–4945. PMLR, 2018.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

A. Related Works

Before diving into the details of the proofs, we provide a more in-depth discussion of previous works that are linked to ours.

In recent years, there has been a tremendous amount of interest in developing and improving policy search algorithms (e.g., Williams, 1992; Baxter & Bartlett, 2001; Lillicrap et al., 2015; Schulman et al., 2015; 2017; Metelli et al., 2018; Cobbe et al., 2021). Most of these methods interleave the two following phases during the training process: first, a batch of trajectories of fixed length is collected by interacting with the environment via Monte Carlo simulation, and, second, these data are used to update the parameters of the policy. Motivated by this setting, we study whether it is possible to optimize Monte Carlo simulations, by exploiting the reset possibility available in many real-world simulators.

To tackle this budget allocation problem, we started with analyzing the problem of estimating, with finite budget Λ, the performance of a given policy via Monte Carlo simulation (Kalos & Whitlock, 2009; Owen, 2013). We started with the on-policy setting by minimizing generalizations of the Höeffding confidence intervals (Boucheron et al., 2003) around the estimated return. Then, we extended the reasoning to the more intricated off-policy problem, through the lens of importance sampling (Hesterberg, 1988; Owen, 2013), and by building on top of the recent confidence intervals of Metelli et al. (2018). We remark that, in this sense, relevant works that are linked to ours can be found in, e.g., Thomas et al. (2015); Thomas & Brunskill (2016); Dai et al. (2020); Mukherjee et al. (2022). However, it has to be noticed that in previous studies, the focus was on different aspects of the estimation problem (e.g., building a policy that minimizes the variance of estimation for the return of a target policy), while we focus purely on the interaction within the environment.

Given our analysis on the estimation problem, we propose to adopt our optimized data collection strategy, together with our optimized confidence intervals, to extend POIS (Metelli et al., 2018), a recent off-policy optimization algorithm that relies on Monte Carlo simulation. More specifically, in Metelli et al. (2018) the authors developed a principled off-policy method that directly models the uncertainty in the update step via controlling (upper bounds on) the variance of importance weights. In this sense, POIS defines an original concept of trust-region that constraints the target policy to be close to the behavioral one, thus limiting high-variance estimations problems that typically affect off-policy methods (Owen, 2013). We notice that this concept of controlling the dissimilarity between the current policy and the next one is at the core of many well-known policy search methods (e.g., Schulman et al., 2015; 2017). Compared to this line of our work, we improve POIS by optimizing the dependence on the number of collected data in $\mathcal{L}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})$. In this sense, we remark that, while we minimized confidence intervals around the empirical off-policy return (i.e., the POIS adaptive trust-region), other choices might better fit other algorithms: extending well-knowns algorithm such as TRPO and PPO with trajectory truncation mechanisms represents an exciting line for future works.

We also notice that several works have considered the problem of optimizing policy search methods by setting hyper-parameters, such as the learning rate (Pirotta et al., 2013), the batch size (Papini et al., 2017), or the amount of policy exploration (Papini et al., 2020), in a theoretically principled way. Compared to this line of work, in this paper, we are considering a *novel type of hyper-parameter*, which is the budget that is spent by the agent to interact with the environment. To this end, we develop a principled theory that leads to collect truncated trajectories. This concept can be related to the recent strand of model-based policy optimization literature that simulates short trajectories in an estimated model of the considered domain to update the parameters of the policy (Janner et al., 2019; Nguyen et al., 2018; Bhatia et al., 2022). Similar ideas on shortened/adaptive horizon have also arisen in the fields of multi-task reinforcement learning (Farahmand et al., 2016) and imitation learning (Sun et al., 2018). However, in all these works, the motivation, the idea, the method, and the analysis completely differ. More specifically, we develop a theory that optimizes the interaction with the environment by exploiting the structure of the RL return. In this sense, our work is complementary to approaches that evolve the discount factor online to form a sort of curriculum (François-Lavet et al., 2015). In particular, we notice that our method could be integrated into these approaches, by evolving the DCS based on the current value of the discount factor. Finally, most recently, Poiani et al. (2022) have introduced the idea of cutting trajectories while interacting with the environment to obtain a biased estimate of the return in planning algorithms such as depth-first search. This sort of idea was presented as an *application* in the context of multi-fidelity bandits (e.g., Kandasamy et al., 2016; 2019), where the crucial trade-off is between the introduced bias (i.e., the maximum error due to cutting the search at a given depth) and the cost of acquiring a given sample (i.e., the number of nodes generated by the algorithm). In this paper, we take a similar perspective, but we develop an approach that is tailored to Monte Carlo Reinforcement Learning settings.

B. Proof and Derivations

In this Section, we provide proofs and derivations for all our theoretical claims. More specifically, Section B.1 provides results for on-policy evaluation, Section B.2 extends these results to the off-policy evaluation setting and Section B.3 provides further theoretical analysis.

B.1. On-Policy Results

We begin by proving that Equation (4) is an unbiased estimate for $J(\boldsymbol{\theta})$.

Theorem B.1. *Consider an optimization budget $\Lambda \geq T$ and a DCS such that $m_T \geq 1$. Consider a policy $\pi_{\boldsymbol{\theta}} \in \Pi_\Theta$. Then:*

$$\mathop{\mathbb{E}}_{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}\left[\hat{J}_{\boldsymbol{m}}(\boldsymbol{\theta})\right] = J(\boldsymbol{\theta}). \tag{13}$$

Proof. Let $r_{t,\boldsymbol{\theta}}$ be the expected t-th reward under policy $\pi_{\boldsymbol{\theta}}$. It is easy to verify that:

$$\mathop{\mathbb{E}}_{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}\left[\hat{J}_{\boldsymbol{m}}(\boldsymbol{\theta})\right] = \sum_{h=1}^{T} m_h \mathop{\mathbb{E}}_{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta},h)}\left[\sum_{t=0}^{h-1}\gamma^t \frac{R(s_t,a_t)}{n_t}\right] = \sum_{h=1}^{T} m_h \sum_{t=0}^{h-1}\gamma^t \frac{r_{t,\boldsymbol{\theta}}}{n_t}$$

At this point, unrolling the summation, we notice that, fixed $\bar{t} \in \{0,\dots,T-1\}$ (i.e., the inner summation), its contribution appears in all $h \in \{1,\dots,T\}$ (i..e, outer summation) such that $h > \bar{t}$. Moreover, since $m_T \geq 1$, all $t \in \{0,\dots,T-1\}$ appears at least once. Therefore:

$$\sum_{h=1}^{T} m_h \sum_{t=0}^{h-1}\gamma^t \frac{r_{t,\boldsymbol{\theta}}}{n_t} = \sum_{t=0}^{T-1}\gamma^t \frac{r_{t,\boldsymbol{\theta}}}{n_t} \sum_{h=t+1}^{T} m_h.$$

However, given the relationship between $\boldsymbol{n}$ and $\boldsymbol{m}$, we have that:

$$\sum_{h=t+1}^{T} m_h = n_t - n_{t+1} + n_{t+1} - n_{t-2} + \dots + n_{T-2} - n_{T-1} + n_{T-1} = n_t.$$

Therefore:

$$\sum_{t=0}^{T-1}\gamma^t \frac{r_{t,\boldsymbol{\theta}}}{n_t} \sum_{h=t+1}^{T} m_h = \sum_{t=0}^{T-1}\gamma^t r_{t,\boldsymbol{\theta}} = J(\boldsymbol{\theta}),$$

which concludes the proof.

□

We now continue with a key technical Lemma that will be crucial to derive our generalizations of the Höeffding confidence intervals.

Lemma B.2. *Consider an arbitrary DCS $\boldsymbol{m}$ such that $m_T \geq 1$. Then:*

$$\sum_{h=1}^{T} m_h \left(\sum_{t=0}^{h-1}\frac{\gamma_t}{n_t}\right)^2 = \sum_{t=0}^{T-1}\frac{c_t}{n_t} \tag{14}$$

where $c_t = \frac{\gamma^t(\gamma^t+\gamma^{t+1}-2\gamma^T)}{1-\gamma}$.

Proof. Consider:

$$\sum_{h=1}^{T} m_h \left(\sum_{t=0}^{h-1}\frac{\gamma^t}{n_t}\right)^2 = \sum_{h=1}^{T} m_h \left(\sum_{t=0}^{h-1}\frac{\gamma^{2t}}{n_t^2} + \sum_{t=0}^{h-2}\sum_{t'=t+1}^{h-1}\frac{2\gamma^{t+t'}}{n_t n_{t'}}\right). \tag{15}$$

Then, focus on the first component, namely, $\sum_{h=1}^{T} m_h \sum_{t=0}^{h-1} \frac{\gamma^{2t}}{n_t^2}$. By unrolling the summations, we notice that, since $m_T \geq 1$, fixed $\bar{t} \in \{0, \dots, T-1\}$ each component $\frac{\gamma^{2t}}{n_t^2}$ appears for all h such that $h > t$. Therefore:

$$\sum_{h=1}^{T} m_h \sum_{t=0}^{h-1} \frac{\gamma^{2t}}{n_t^2} = \sum_{t=0}^{T-1} \left(\frac{\gamma^{2t}}{n_t^2} \sum_{h=t+1}^{T} m_h \right) = \sum_{t=0}^{T-1} \frac{\gamma^{2t}}{n_t}, \tag{16}$$

where in the last passage, we have used $\sum_{h=t+1}^{T} m_h = n_t$, which directly follow by the relationship between $\boldsymbol{n}$ and $\boldsymbol{m}$ and the fact that $m_T \geq 1$.

Then, consider the second part of Equation (15), namely $\sum_{h=1}^{T} m_h \left(\sum_{t=0}^{h-2} \sum_{t'=t+1}^{h-1} \frac{2\gamma^{t+t'}}{n_t n_{t'}} \right)$. By unrolling the summations, we notice that fixed the outer index of the inner summation, i.e., $t = \bar{t} \in \{0, \dots, T-2\}$, its contribution will appear only for $h > t+1$, thus leading to:

$$\sum_{h=1}^{T} m_h \left(\sum_{t=0}^{h-2} \sum_{t'=t+1}^{h-1} \frac{2\gamma^{t+t'}}{n_t n_{t'}} \right) = \sum_{t=0}^{T-2} \sum_{h=t+2}^{T} m_h \sum_{t'=t+1}^{h-1} \frac{2\gamma^{t+t'}}{n_t n_{t'}}.$$

At this point, fix $\bar{t} \in \{0, \dots, T-2\}$ and consider $\sum_{h=\bar{t}+2}^{T} m_h \sum_{t'=\bar{t}+1}^{h-1} \frac{2\gamma^{\bar{t}+t'}}{n_{\bar{t}} n_{t'}}$. Unrolling the summation, we notice that a given term t' appears only for $h > t'$. Therefore:

$$\sum_{h=\bar{t}+2}^{T} m_h \sum_{t'=\bar{t}+1}^{h-1} \frac{2\gamma^{\bar{t}+t'}}{n_{\bar{t}} n_t'} = \sum_{t'=\bar{t}+1}^{T-1} \frac{2\gamma^{\bar{t}+t'}}{n_{\bar{t}} n_{t'}} \left(\sum_{h=t'+1}^{T} m_h \right) = \sum_{t'=\bar{t}+1}^{T-1} \frac{2\gamma^{\bar{t}+t'}}{n_{\bar{t}}}. \tag{17}$$

Using Equations (16) and (17) in Equation (15), we have that:

$$\begin{aligned}
\sum_{h=1}^{T} m_h \left(\sum_{t=0}^{h-1} \frac{\gamma^t}{n_t} \right)^2 &= \sum_{t=0}^{T-1} \frac{\gamma^{2t}}{n_t} + \sum_{t=0}^{T-2} \sum_{t'=t+1}^{T-1} \frac{2\gamma^{t'+t}}{n_t} \\
&= \sum_{t=0}^{T-1} \frac{\gamma^{2t}}{n_t} + \sum_{t=0}^{T-2} \frac{2\gamma^t}{n_t} \sum_{t'=t+1}^{T-1} \gamma^{t'} \\
&= \sum_{t=0}^{T-1} \frac{\gamma^{2t}}{n_t} + \sum_{t=0}^{T-2} \frac{2\gamma^t}{n_t} \left(\frac{\gamma^{t+1} - \gamma^T}{1-\gamma} \right) \\
&= \sum_{t=0}^{T-1} \frac{\gamma^{2t}}{n_t} + \sum_{t=0}^{T-1} \frac{2\gamma^t}{n_t} \left(\frac{\gamma^{t+1} - \gamma^T}{1-\gamma} \right) \\
&= \sum_{t=0}^{T-1} \frac{\gamma^t(\gamma^t + \gamma^{t+1} - 2\gamma^T)}{1-\gamma} \cdot \frac{1}{n_t} \\
&= \sum_{t=0}^{T-1} \frac{c_t}{n_t},
\end{aligned}$$

which concludes the proof. □

At this point, we are ready to prove Theorem 3.2. We first report, for completeness, the Höeffding's inequality for the sum of subgaussian random variables.

Lemma B.3. *Let $X_1, \dots, X_n$ be independent sub-gaussian r.v. with mean $\mu_1, \dots, \mu_n$ and subgaussianity parameters $\sigma_1^2, \dots, \sigma_n^2$, respectively. Let $\bar{\mu}_n \coloneqq \sum_{i=1}^n \mu_i$ and $\hat{\mu}_n \coloneqq \sum_{i=1}^n X_i$. Then, $\forall \epsilon > 0$, it holds that:*

$$\mathbb{P}(|\hat{\mu}_n - \bar{\mu}_n| > \epsilon) \leq 2 \exp\left(-\frac{\epsilon^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Proposition B.4. *Consider an optimization budget $\Lambda \geq T$, a generic DCS $\boldsymbol{m}$ such that $m_T \geq 1$ and $\delta \in (0,1)$. Then, with probability at least $1-\delta$ it holds that:*

$$\left|\hat{J}_{\boldsymbol{m}}(\boldsymbol{\theta}) - J(\boldsymbol{\theta})\right| \leq \sqrt{\frac{1}{2}\log\left(\frac{2}{\delta}\right)\sum_{t=0}^{T-1}\frac{c_t}{n_t}}, \tag{5}$$

where $c_t = \frac{\gamma^t(\gamma^t + \gamma^{t+1} - 2\gamma^T)}{1-\gamma}$.

Proof. First of all, we notice that rewards are bounded in $[0,1]$. It follows that, given a trajectory of length h, $\sum_{t=0}^{h-1}\gamma^t \frac{R(s_t,a_t)}{n_t}$ is a subgaussian r.v. with subgaussianity parameter $\sigma_h^2 = \frac{1}{4}\left(\sum_{t=0}^{h-1}\frac{\gamma^t}{n_t}\right)^2$.[10] It follows that we can treat $\hat{J}_{\boldsymbol{m}}(\boldsymbol{\theta})$ as a sum of random variables with expected value $J(\boldsymbol{\theta})$. Therefore, we can apply Lemma B.3 with $\epsilon = \sqrt{2\sum_{h=1}^T m_h \sigma_h^2 \log(2/\delta)}$, obtaining that, with probability at least $1-\delta$:

$$|J_{\boldsymbol{m}}(\boldsymbol{\theta}) - J(\boldsymbol{\theta})| \leq \sqrt{2\sum_{h=1}^T m_h \sigma_h^2 \log(2/\delta)} = \sqrt{\frac{1}{2}\sum_{h=1}^T m_h \left(\sum_{t=0}^{h-1}\frac{\gamma^t}{n_t}\right)^2 \log(2/\delta)}$$

The result, then, follow by combining the previous Equation with Lemma B.2. □

At this point, our focus shifts toward finding our approximately optimal DCS $\tilde{\boldsymbol{m}}^*$. We will derive our results for the general any-budget case, after which Theorem 3.3 will follow as a special case. Our proofs follow by combining a closed form solution of a relaxation of (6), where we drop the integer constraints on n_t, and some integrality gap arguments. More specifically, we are interested in the following relaxation of (6):

$$\begin{aligned} \min_{\boldsymbol{n}} \quad & \sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{n_t}} \\ \text{s.t.} \quad & \sum_{t=0}^{T-1} n_t = \Lambda \\ & n_t \geq n_{t+1}, \quad \forall t \in \{0, \dots, T-2\} \\ & n_t \geq 1, \quad \forall t \in \{0, \dots, T-1\} \end{aligned} \tag{18}$$

where the only difference stands in the fact that $n_t \in \mathbb{N}_+$ has now been replaced with $n_t \geq 1$. For this reason, we first present a simplified version of (18), that preserves the optimal solution.

Lemma B.5. *Consider an optimization $\Lambda \geq T$. The convex relaxation of the optimization problem* (6) *can be written as:*

$$\begin{aligned} \min_{\boldsymbol{n}} \quad & \sum_{t=0}^{T-1}\frac{c_t}{n_t} \\ \textit{s.t.} \quad & \sum_{t=0}^{T-1} n_t = \Lambda \\ & n_t \geq 1, \quad \forall t \in \{0, \dots, T-1\} \end{aligned} \tag{19}$$

where $\boldsymbol{n} = (n_0, \dots, n_{T-1})$ and $c_t = \gamma^t(\gamma^t + \gamma^{t+1} - 2\gamma^T)$ Furthermore, the optimization problem (19) *is convex in $\boldsymbol{n}$.*

[10] We recall that a r.v. with bounded support over $[a,b]$ $(b > a)$ is sub-gaussian with scale given by $\frac{(b-a)^2}{4}$

Proof. First, we prove the equivalence of the objective function. We notice that since the square root is a monotic function, it does not affect the optimal solution. Moreover, $\log(2/\delta)$ can be seen as a constant and, therefore, it can be neglected from the objective function as well. Thus, the optimal solution of the problem is preserved.

Then, we prove the equivalence of the constraints. The only difference between (19) and (18) lies in the fact that $n_t \geq n_{t+1}$ has been neglected from the formulation. Given the structure of the simplified objective function (i.e., $\sum_{t=0}^{T-1} \frac{c_t}{n_t}$), we notice that $n_t \geq n_{t+1}$ will always be satisfied for an optimal solution. Indeed, suppose that $n_t < n_{t+1}$ for some t; then, since $c_t > c_{t+1}$, we can always improve the value of the objective function by swapping n_t with n_{t+1}. Therefore, it is possible to neglect these constraints given that $n_t \geq 1$.

Finally, we conclude by proving the convexity of (19). First of all, we notice that all constraints are linear (and, thus, convex). It remains to prove the convexity of the objective function, i.e., $\sum_{t=0}^{T-1} \frac{c_t}{n_t}$. We begin by remarking that $\sum_{t=0}^{T-1} \frac{c_t}{n_t}$ is infinitely differentiable over the domain $(\mathbb{R} - \{0\})^T$. In this case, to prove the convexity it is sufficient to ensure that the Hessian matrix is positive semidefinite. More specifically, in our case, the Hessian matrix is a diagonal matrix where the i-th element of the diagonal is given by $\frac{2c_i}{n_i^3}$, which, thus, is positive definite. This concludes the proof. □

At this point, we derive a closed-form solution for (18) by analyzing the KKT condition of (19) (Boyd et al., 2004).

Lemma B.6. *Consider an optimization budget $\Lambda > T$, and consider $h \in \{1, \dots, T\}$. Let $n_t(h) = 1$ for $t \geq h$ and $n_t(h) = \frac{\sqrt{c_t}}{\sum_{i=0}^{h-1} \sqrt{c_i}}(\Lambda - T + h)$ for $t < h$. The optimal value of the objective function of the convex relaxation of* (6) *can be computed as:*

$$\min_{h \in \{1,\dots,T\}} \sqrt{\frac{1}{2} \log(2/\delta) \sum_{t=0}^{T-1} \frac{c_t}{n_t(h)}} \tag{20}$$

and h is such that, for all $t \geq h$ it holds that:

$$\Lambda - T + h \leq \frac{\sum_{i=0}^{h-1} \sqrt{c_i}}{\sqrt{c_t}}, \tag{21}$$

and, for all $t < h$:

$$\Lambda - T + h > \frac{\sum_{i=0}^{h-1} \sqrt{c_i}}{\sqrt{c_t}}. \tag{22}$$

Proof. Due to Lemma B.5, we can study the optimal value of the convex relaxation by analyzing (19). More specifically, we focus on the following variant:

$$\begin{aligned} \min_{\boldsymbol{n}} \quad & \sum_{t=0}^{T-1} \frac{c_t}{\tilde{n}_t + 1} \\ \text{s.t.} \quad & \tilde{n}_t \geq 0 \quad \forall t \in \{0, \dots, T-1\} \\ & \sum_{t=0}^{T-1} \tilde{n}_t \leq \Lambda' \end{aligned} \tag{23}$$

where $\Lambda' = \Lambda - T$, and the fact that $n_t \geq 1$ has been directly forced in the objective function and in the constraints by applying the change of variables $n_t = \tilde{n}_t + 1$.

At this point, the KKT conditions for the optimization problem (23) are given by:

$$\begin{cases} -\frac{c_t}{(\tilde{n}_t+1)^2} - \mu_t + \eta = 0 & \forall t \in \{0, \dots, T-1\} \\ \mu_t \tilde{n}_t = 0 & \forall t \in \{0, \dots, T-1\} \\ \eta(\sum_{t=0}^{T-1} \tilde{n}_t - \Lambda') = 0 & \\ \sum_{t=0}^{T-1} \tilde{n}_t - \Lambda' \leq 0 & \\ \mu_t \geq 0 & \forall t \in \{0, \dots, T-1\} \\ \eta \geq 0 & \end{cases}. \tag{24}$$

Since the problem is convex, the solution to the KKT conditions are the global optimum of the problem. To find it, we begin with the first equation for a general $t \in \{0, \dots, T-1\}$. From algebraic manipulations we obtain:

$$\tilde{n}_t = \sqrt{\frac{c_t}{\eta - \mu_t}} - 1. \tag{25}$$

At this point, we split our analysis into two cases that arise from the second equation of the system (24). More specifically, we notice that when $n_t > 0$, the second equation of the system (24) leads to $\mu_t = 0$, [11] which reduces Equation (25) to:

$$\tilde{n}_t = \sqrt{\frac{c_t}{\eta}} - 1. \tag{26}$$

Therefore, since $c_t > 0$, this implies $\eta > 0$. At this point, since $\eta > 0$, the third equation of the system (24) leads to:

$$\sum_{t:\tilde{n}_t > 0} \tilde{n}_t = \sum_{t:\tilde{n}_t > 0} \left(\sqrt{\frac{c_t}{\eta}} - 1 \right) = \Lambda'. \tag{27}$$

However, due to the structure of the objective function, for $i, j \in \{0, \dots, T-1\}$ such that $i > j$, if $\tilde{n}_i > 0$, then $\tilde{n}_j > 0$. The main intuition is that, otherwise, we could set $\tilde{n}_j$ to the value of $\tilde{n}_i$, and $\tilde{n}_i$ to 0, and improve the objective function.[12]

Therefore, we can write Equation (27) as a general function of an integer $h \in \{1, \dots, T\}$ that indicates the first time-step t for which $n_h = 0$ holds.[13] More specifically, Equation (27) reduces to:

$$\sum_{t=0}^{h-1} \left(\sqrt{\frac{c_t}{\eta}} - 1 \right) = \Lambda'. \tag{28}$$

Solving Equation (28) for η, we obtain:

$$\eta = \left(\frac{\sum_{i=0}^{h-1} \sqrt{c_i}}{\Lambda' + h} \right)^2, \tag{29}$$

which, as we can appreciate is always greater than 0, thus satisfying the constraints imposed so far.

At this point, using Equation (29) in Equation (26), we obtain:

$$\tilde{n}_t = \frac{\sqrt{c_t}}{\sum_{i=0}^{h-1} \sqrt{c_i}} (\Lambda' + h) - 1, \tag{30}$$

[11] We notice that this satisfies the last constraint $\mu_t \geq 0$.

[12] This follows from the fact that $c_i < c_j$.

[13] We notice that, since $\Lambda' > 0$, we always have at least $\tilde{n}_0 > 0$.

which holds for a generic $t < h$, under the constraint that:

$$\frac{\sqrt{c_t}}{\sum_{i=0}^{h-1}\sqrt{c_i}}(\Lambda' + h) - 1 > 0.$$

Or, equivalently:

$$\Lambda' + h > \frac{\sum_{i=0}^{h-1}\sqrt{c_i}}{\sqrt{c_t}}. \tag{31}$$

Now, we consider the second case (i.e., $\tilde{n}_t = 0$) that arises from the second equation of the system (24). In this case, $\tilde{n}_t = 0$ and μ_t is possibly different from 0, and we need to ensure that $\mu_t \geq 0$ to satisfy the last constraint of (24). More specifically, this reduces to study:

$$-c_t - \mu_t + \left(\frac{\sum_{i=0}^{h-1}\sqrt{c_i}}{\Lambda' + h}\right)^2 = 0,$$

for $h \geq t$. In particular, we obtain:

$$\mu_t = \left(\frac{\sum_{i=0}^{h-1}\sqrt{c_i}}{\Lambda' + h}\right)^2 - c_t,$$

which we need to impose as greater or equal than 0, that is:

$$\left(\frac{\sum_{i=0}^{h-1}\sqrt{c_i}}{\Lambda' + h}\right)^2 - c_t.$$

Or, equivalently:

$$\Lambda' + h \leq \frac{\sum_{i=0}^{h-1}\sqrt{c_i}}{\sqrt{c_t}}. \tag{32}$$

At this point, we remark that there always at least exists one value of h for which both Equations (31) and (32) are satisfied. Indeed, due to the Weierstrass theorem, if the objective function is continuous and considered on a closed and bounded domain, then a global optimum exists. Moreover, the KKT are sufficient conditions for global optimality in convex problems, from which follows the existence of at least one h satisfying both equations.

Putting everything together, and rescaling $\tilde{n}_t$ to n_t concludes the proof.

□

Lemma B.7. *Consider an optimization budget $\Lambda > T$, and consider $h \in \{1, \dots, T\}$. Let $n_t(h) = 1$ for $t \geq h$ and $n_t(h) = \frac{\sqrt{c_t}}{\sum_{i=0}^{h-1}\sqrt{c_i}}(\Lambda - T + h)$ for $t < h$. The optimal value of the objective function of the optimization problem* (18) *can be computed as:*

$$\sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{n_t(h^*)}} \tag{33}$$

where h^* is the only $h \in \{1, \dots, T\}$ for which the following holds: for all $t \geq h^*$:

$$\Lambda - T + h^* \leq \frac{\sum_{i=0}^{h^*-1} \sqrt{c_i}}{\sqrt{c_t}} \tag{34}$$

and, for all $t < h^*$:

$$\Lambda - T + h^* > \frac{\sum_{i=0}^{h^*-1} \sqrt{c_i}}{\sqrt{c_t}}. \tag{35}$$

Proof. From Lemma B.6, what remains to prove is that there exists a single $h \in \{1, \dots, T\}$ for which Equations (34) and (35) are satisfied. Since at least one h exists, we need to prove that it is impossible that Equations (34) and (35) are satisfied for two distinct $\bar{h}_1, \bar{h}_2 \in \{1, \dots, T\}$. Suppose w.l.o.g. that $\bar{h}_1 > \bar{h}_2$ and proceed by contradiction.

First of all, consider a generic h and focus on Equation (34). A sufficient condition for Equation (34) to hold for all $t \geq h$ is given by:

$$\Lambda - T + h \leq \frac{\sum_{i=0}^{h-1} \sqrt{c_i}}{\sqrt{c_h}}. \tag{36}$$

Similarly, a sufficient condition for Equation (35) to hold for all $t < h$ is given by:

$$\Lambda - T + h > \frac{\sum_{i=0}^{h-1} \sqrt{c_i}}{\sqrt{c_{h-1}}}. \tag{37}$$

Now, consider Equation (36) for $\bar{h}_2$ and Equation (37) for $\bar{h}_1$, namely:

$$\begin{cases} \Lambda - T + \bar{h}_2 \leq \frac{\sum_{i=0}^{\bar{h}_2-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_2}}} \\ \frac{\sum_{i=0}^{\bar{h}_1-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_1-1}}} < \Lambda - T + \bar{h}_1 \end{cases}. \tag{38}$$

Summing the two equations of System (38), and rearranging the terms we obtain:

$$\bar{h}_1 - \bar{h}_2 > \frac{\sum_{i=0}^{\bar{h}_1-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_1-1}}} - \frac{\sum_{i=0}^{\bar{h}_2-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_2}}}.$$

However, as we shall show in a moment:

$$\bar{h}_1 - \bar{h}_2 \leq \frac{\sum_{i=0}^{\bar{h}_1-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_1-1}}} - \frac{\sum_{i=0}^{\bar{h}_2-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_2}}}, \tag{39}$$

always holds, thus leading to a contradiction. Indeed, consider:

$$\frac{\sum_{i=0}^{\bar{h}_1-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_1-1}}} - \frac{\sum_{i=0}^{\bar{h}_2-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_2}}} = \frac{\sum_{i=\bar{h}_2}^{\bar{h}_1-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_1-1}}} + \frac{\sum_{i=0}^{\bar{h}_2-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_1-1}}} - \frac{\sum_{i=0}^{\bar{h}_2-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_2}}}.$$

However,

$$\frac{\sum_{i=\bar{h}_2}^{\bar{h}_1-1} \sqrt{c_i}}{\sqrt{c_{\bar{h}_1-1}}} \geq \frac{\sum_{i=\bar{h}_2}^{\bar{h}_1-1} \sqrt{c_{\bar{h}_1-1}}}{\sqrt{c_{\bar{h}_1-1}}} = \bar{h}_1 - \bar{h}_2.$$

Moreover:

$$\frac{\sum_{i=0}^{\bar{h}_2-1}\sqrt{c_i}}{\sqrt{c_{\bar{h}_1-1}}}-\frac{\sum_{i=0}^{\bar{h}_2-1}\sqrt{c_i}}{\sqrt{c_{\bar{h}_2}}}\geq\frac{\sum_{i=0}^{\bar{h}_2-1}\sqrt{c_i}}{\sqrt{c_{\bar{h}_2}}}-\frac{\sum_{i=0}^{\bar{h}_2-1}\sqrt{c_i}}{\sqrt{c_{\bar{h}_2}}}=0.$$

From which it follows Equation (39), thus concluding the proof.

□

Lemma B.7 deserves some comments. First of all, it provides an $\mathcal{O}(T)$ procedure to compute the optimal solution of the convex relaxation of (6). Indeed, it is sufficient to iterate over the variable h to find the only h^* for which Equations (34) and (35) holds. In the rest of this text, we will refer to the optimal solution of the convex relaxation as $\bar{\boldsymbol{n}}^*$. Secondly, we notice that Equations (34) and (35) put in a tight relationship $\bar{\boldsymbol{n}}^*$ with the available budget Λ. In particular, when the budget is sufficiently small (i.e., $\Lambda - T + h^* \leq \frac{\sum_{i=0}^{h^*-1}\sqrt{c_i}}{\sqrt{c_t}}$), $\bar{\boldsymbol{n}}^*$ will necessarily allocate a single sample (i.e., $n_t(h^*)=1$) for all $t \geq h^*$, which corresponds to the minimal amount of samples that is required to obtain an unbiased estimate. Conversely, when the budget is sufficiently large (i.e., $\Lambda - T + h^* > \frac{\sum_{i=0}^{h^*-1}\sqrt{c_i}}{\sqrt{c_t}}$), $\bar{\boldsymbol{n}}^*$ will allocate to all $t < h^*$ a non-uniform budget quantity that is given by $\frac{\sqrt{c_t}}{\sum_{i=0}^{h^*-1}\sqrt{c_i}}(\Lambda - T + h^*)$.

At this point, we notice that, to study the solution in its closed form, we relaxed integer constraints. Given the relaxed solution of Lemma B.7, it is easy to obtain a proper DCS by taking the element-wise floor of the optimal relaxed solution and allocating the remaining budget uniformly. The resulting DCS, which we refer to as the approximately optimal DCS $\tilde{\boldsymbol{n}}^*$, will thus differ at most by 1 (element-wise) w.r.t. to the optimal relaxed solution of Lemma B.7. More formally, we define the any-budget $\tilde{\boldsymbol{n}}^*$ in the followin way.

Definition B.8. Consider an optimization budget $\Lambda > T$ and consider $h \in \{1, \dots, T\}$. Let $n_t(h) = 1$ for $t \geq h$ and $n_t(h) = \frac{\sqrt{c_t}}{\sum_{i=0}^{h-1}\sqrt{c_i}}(\Lambda - T + h)$ for $t < h$. Let h^* be the only h such that Equation (34) and Equation (35) are satisfied. Let $k = \Lambda - \sum_{t=0}^{T-1}\lfloor n_t(h^*)\rfloor$. Then, we define the approximately optimal DCS $\tilde{\boldsymbol{n}}^* = (\tilde{n}_0^*, \dots, \tilde{n}_{T-1}^*)$, where:

$$\tilde{n}_t^* = \lfloor n_t(h^*)\rfloor + \mathbf{1}\{t < k\} \tag{40}$$

Notice that Definition B.8 reduces to the one of Theorem 3.3 for sufficiently large budget. More specifically, taking $\Lambda \geq \Lambda_0 = \frac{\sum_{t=0}^{T-1}\sqrt{c_t}}{\sqrt{c_{T-1}}}$, then $h^* = T$, from which follows the expression of Theorem 3.3.

Lemma B.9. *Consider an optimization budget $\Lambda > T$, let $\bar{\boldsymbol{n}}$ be the optimal solution of optimal solution of the convex relaxation given in Lemma B.7, and let $\tilde{\boldsymbol{n}}^*$ as in Defintion B.8. Then, for each $t \in \{0, \dots, T-1\}$, $|\bar{n}_t - \tilde{n}_t^*| \leq 1$.*

Proof. Consider t such that $t \geq k$ holds. Then $|\bar{n}_t - \tilde{n}_t^*| = |n_t(h^*) - \lfloor n_t(h^*)\rfloor| \leq 1$.

Consider t such that $t < k$ holds. Then $|\bar{n}_t - \tilde{n}_t^*| = |n_t(h^*) - \lfloor n_t(h^*)\rfloor + 1| \leq 1$, which concludes the proof. □

At this point, what is left is analyzing the quality of the approximately optimal DCS $\tilde{\boldsymbol{n}}^*$, which will to the proof of Theorem 3.3. We begin by reporting the equivalent version of Theorem 3.3 that holds for the generic case of $\Lambda > T$.[14]

Theorem B.10. *Consider an optimization budget $\Lambda > T$, let $\tilde{\boldsymbol{n}}^*$ be the approximately optimal DCS given in Defintion B.8, and let $\boldsymbol{n}^*$ be the optimal solution of the integer optimization problem* (6)*. Moreover, let $f(\boldsymbol{n}) = \sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{n_t}}$. Then,*

$$f(\boldsymbol{n}^*) \leq f(\tilde{\boldsymbol{n}}^*) \leq \sqrt{2} f(\boldsymbol{n}^*) \tag{41}$$

Proof. First of all, focus $1 \leq \frac{f(\tilde{\boldsymbol{n}}^*)}{f(\boldsymbol{n}^*)}$. This clearly holds since $\boldsymbol{n}^*$ is the optimal solution of (6), while $\tilde{\boldsymbol{n}}^*$ is a feasibile solution.

[14] Notice that for $\Lambda = T$ there exists only one DCS that satisfies $m_T \geq 1$, and, consequently, the problem is trivial.

Now, what remains to prove is that $\frac{f(\tilde{\boldsymbol{n}}^*)}{f(\boldsymbol{n}^*)} \leq \sqrt{2}$. Let $\bar{\boldsymbol{n}}^*$ be the optimal solution of the convex relaxation given in Lemma B.7. Then, first of all, we notice that:

$$\frac{f(\tilde{\boldsymbol{n}}^*)}{f(\boldsymbol{n}^*)} \leq \frac{f(\tilde{\boldsymbol{n}}^*)}{f(\bar{\boldsymbol{n}}^*)} \tag{42}$$

holds since $\bar{\boldsymbol{n}}^*$ is the optimal solution of the same optimization problem but with a removed constraint (i.e., the integer constraint on n_t). Then, consider:

$$f(\bar{\boldsymbol{n}}^*) = \sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{\bar{n}_t^*}} \geq \sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{\tilde{n}_t^*+1}} \geq \sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{2\tilde{n}_t^*}} = \sqrt{\frac{1}{2}}f(\tilde{\boldsymbol{n}}^*) \tag{43}$$

where the first inequality follows from Lemma B.9 and the second one by $\tilde{n}_t^* \geq 1$. Plugging Equation (43) into Equation (42) concludes the proof. □

Theorem 3.3. *Consider an optimization budget $\Lambda \geq \Lambda_0$, let $\boldsymbol{n}^*$ be the optimal solution of* (6)*. Let $g_t = \frac{\sqrt{c_t}}{\sum_{i=0}^{T-1}\sqrt{c_i}}\Lambda$, and let $k = \Lambda - \sum_{t=0}^{T-1}\lfloor g_t \rfloor$. Define the t-th component of the approximately optimal DCS $\tilde{\boldsymbol{n}}^*$ as $\tilde{n}_t^* \coloneqq \lfloor g_t \rfloor + \mathbf{1}\{t < k\}$. Then, it holds that:*

$$f(\boldsymbol{n}^*) \leq f(\tilde{\boldsymbol{n}}^*) \leq \sqrt{2}f(\boldsymbol{n}^*). \tag{7}$$

Proof. The proof is a direct consequence of Theorem B.10. □

We now continue by providing the PAC analysis for our approximately optimal DCS $\tilde{\boldsymbol{n}}^*$. Before diving into the proof of Theorem 3.4, we provide two intermediate technical results.

Lemma B.11. *Consider an optimization budget $\Lambda \geq 2T$ and let $\tilde{\boldsymbol{n}}^*$ be as in Definition B.8. Then:*

$$\sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{\tilde{n}_t^*}} \leq \sqrt{2\frac{\log(2/\delta)}{\Lambda}\left(\sum_{t=0}^{T-1}\sqrt{c_t}\right)^2} \tag{44}$$

Proof. Let $\bar{\boldsymbol{n}}^*$ be the optimal solution of the convex relaxation given in Lemma B.7. Then, as in Theorem B.10, we have that:

$$\sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{\tilde{n}_t^*}} \leq \sqrt{\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{\bar{n}_t^*}}$$

Now, plugging in the definition of $\bar{\boldsymbol{n}}^*$, we have that:

$$\sqrt{\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{\bar{n}_t^*}} = \sqrt{\log(2/\delta)\sum_{t=0}^{h^*-1}\frac{c_t}{\bar{n}_t^*} + \log(2/\delta)\sum_{t=h^*}^{T-1}c_t} \tag{45}$$

$$= \sqrt{\log(2/\delta)\sum_{t=0}^{h^*-1}\frac{c_t}{\sqrt{c_t}(\Lambda - T + h^*)}\sum_{i=0}^{h^*-1}\sqrt{c_i} + \log(2/\delta)\sum_{t=h^*}^{T-1}c_t} \tag{46}$$

where h^* is the only h that satisfies Equation (34) and (35).

Now, since $\Lambda \geq 2T$, we have that:

$$\sum_{t=0}^{h^*-1}\frac{c_t}{\sqrt{c_t}(\Lambda - T + h^*)}\sum_{i=0}^{h^*-1}\sqrt{c_i} \leq \sum_{t=0}^{h^*-1}\frac{c_t}{\sqrt{c_t}(\Lambda - 1/2\Lambda)}\sum_{i=0}^{h^*-1}\sqrt{c_i} \tag{47}$$

$$\leq 2\sum_{t=0}^{h^*-1}\frac{c_t}{\sqrt{c_t}\Lambda}\sum_{i=0}^{h^*-1}\sqrt{c_i} \tag{48}$$

$$\leq \frac{2}{\Lambda}\sum_{t=0}^{h^*-1}\sqrt{c_t}\sum_{i=0}^{T-1}\sqrt{c_i} \tag{49}$$

which, plugged into Equation (45), leads to:

$$\sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{\tilde{n}_t^*}} \leq \sqrt{\log(2/\delta)\frac{2}{\Lambda}\sum_{t=0}^{h^*-1}\sqrt{c_t}\sum_{i=0}^{T-1}\sqrt{c_i} + \log(2/\delta)\sum_{t=h^*}^{T-1}c_t} \tag{50}$$

At this point, focus on $\sum_{t=h^*}^{T-1} c_t$. From Equation (34), we know that for h^* it holds that:

$$\frac{(\Lambda - T + h^*)\sqrt{c_t}}{\sum_{i=0}^{h^*-1}\sqrt{c_i}} \leq 1$$

Therefore,

$$\sum_{t=h^*}^{T-1} c_t = \sum_{t=h^*}^{T-1}\frac{c_t}{1} \leq \sum_{t=h^*}^{T-1}\frac{c_t}{\sqrt{c_t}(\Lambda - T + h^*)}\sum_{i=0}^{h^*-1}\sqrt{c_i} \leq \frac{2}{\Lambda}\sum_{t=h^*}^{T-1}\sqrt{c_t}\sum_{i=0}^{T-1}\sqrt{c_i} \tag{51}$$

where in the last inequality we have used the same arguments of Equation (47).

At this point, plugging Equation (51) into Equation (50), leads to:

$$\begin{aligned}\sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{\tilde{n}_t^*}} &\leq \sqrt{\log(2/\delta)\frac{2}{\Lambda}\sum_{t=0}^{h^*-1}\sqrt{c_t}\sum_{i=0}^{T-1}\sqrt{c_i} + \log(2/\delta)\frac{2}{\Lambda}\sum_{t=h^*}^{T-1}\sqrt{c_t}\sum_{i=0}^{T-1}\sqrt{c_i}} \\ &= \sqrt{\frac{2\log(2/\delta)}{\Lambda}\left(\sum_{t=0}^{T-1}\sqrt{c_t}\right)^2}\end{aligned}$$

which concludes the proof. □

Lemma B.12. *Consider $\delta \in (0,1)$ and $\epsilon > 0$ such that $\log(2/\delta)c_0 \geq 8T\epsilon^2$ holds. Then $\Lambda \geq 2T$ is a necessary condition to guarantee that $|\hat{J}_{\tilde{\boldsymbol{m}}^*}(\boldsymbol{\theta}) - J(\boldsymbol{\theta})| \leq \epsilon$ holds.*

Proof. We proceed by contradiction: suppose that $\Lambda < 2T$. Then, we continue by lower-bounding the value of the confidence intervals when using our approximately optimal DCS $\tilde{\boldsymbol{m}}^*$. More specifically, let $\bar{\boldsymbol{n}}^*$ be the optimal solution of the convex relaxation of (6). Then, we have that:

$$\sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{\tilde{n}_t^*}} \geq \sqrt{\frac{1}{2}\log(2/\delta)\frac{c_0}{\tilde{n}_0^*}} \geq \sqrt{\frac{1}{2}\log(2/\delta)\frac{c_0}{\bar{n}_0^*+1}} \geq \sqrt{\frac{1}{4}\log(2/\delta)\frac{c_0}{\bar{n}_0^*}}$$

where in the first inequality we have removed positive terms, in the second one we have used Lemma B.9 and in the third one we have used $\bar{n}_0^* \geq 1$. At this point, by noticing that $\bar{n}_0^* < \Lambda$, we obtain:

$$\sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{\tilde{n}_t^*}} \geq \sqrt{\frac{1}{4}\log(2/\delta)\frac{c_0}{\Lambda}}$$

We can now focus on:

$$\sqrt{\frac{1}{4}\log(2/\delta)\frac{c_0}{\Lambda}} \leq \epsilon$$

which, in turn, leads to:

$$\Lambda \geq \frac{1}{4}\log(2/\delta)\frac{c_0}{\epsilon^2}$$

which however, leads to $\Lambda \geq 2T$, thus concluding the proof. □

We are now ready to prove the PAC bound on our approximately optimal DCS $\tilde{\boldsymbol{m}}^*$.

Theorem 3.4. *Let $\delta \in (0,1)$ and $\epsilon > 0$ such that $8T\epsilon^2 \leq \log(2/\delta)c_0$ holds. Then, with probability at least $1-\delta$, $|\hat{J}_{\tilde{\boldsymbol{m}}^*}(\boldsymbol{\theta}) - J(\boldsymbol{\theta})| \leq \epsilon$ holds provided that:*

$$\Lambda = \mathcal{O}\left(\min\left\{\frac{T\log(2/\delta)}{(1-\gamma)^2\epsilon^2}, \frac{\log(2/\delta)}{(1-\gamma)^3\epsilon^2}\right\}\right) \tag{8}$$

Proof. First of all, consider the value of the confidence intervals of $\tilde{\boldsymbol{m}}^*$. Due to Lemma B.12, we know that $\Lambda \geq 2T$ holds; therefore, by applying Lemma B.11 we obtain that:

$$\sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{\tilde{n}_t^*}} \leq \sqrt{2\frac{\log(2/\delta)}{\Lambda}\left(\sum_{t=0}^{T-1}\sqrt{c_t}\right)^2} \leq \epsilon$$

This, in turn, leads to:

$$2\frac{\log(2/\delta)}{\Lambda}\left(\sum_{t=0}^{T-1}\sqrt{c_t}\right)^2 \leq \epsilon^2 \tag{52}$$

At this point, focus on $\left(\sum_{t=0}^{T-1}\sqrt{c_t}\right)^2$:

$$\begin{aligned}\left(\sum_{t=0}^{T-1}\sqrt{c_t}\right)^2 &= \frac{1}{1-\gamma}\left(\sum_{t=0}^{T-1}\sqrt{\gamma^t(\gamma^t+\gamma^{t+1}-2\gamma^T)}\right)^2 \\ &= \frac{1}{1-\gamma}\left(\sum_{t=0}^{T-1}\gamma^t(\gamma^t+\gamma^{t+1}-2\gamma^T) + 2\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\sqrt{\gamma^t(\gamma^t+\gamma^{t+1}-2\gamma^T)}\sqrt{\gamma^{t'}(\gamma^{t'}+\gamma^{t'+1}-2\gamma^T)}\right)\end{aligned}$$

First, consider $\sum_{t=0}^{T-1}\gamma^t(\gamma^t+\gamma^{t+1}-2\gamma^T)$:

$$\sum_{t=0}^{T-1}\gamma^t(\gamma^t+\gamma^{t+1}-2\gamma^T) \leq \sum_{t=0}^{T-1}\gamma^t(\gamma^t+\gamma^{t+1}) \leq \sum_{t=0}^{T-1}\gamma^t(2\gamma^t) \leq 2\sum_{t=0}^{T-1}\gamma^{2t} \leq 2\frac{1-\gamma^T}{1-\gamma}$$

Then, consider $2\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\sqrt{\gamma^t(\gamma^t+\gamma^{t+1}-2\gamma^T)}\sqrt{\gamma^{t'}(\gamma^{t'}+\gamma^{t'+1}-2\gamma^T)}$:

$$\begin{aligned}2\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\sqrt{\gamma^t(\gamma^t+\gamma^{t+1}-2\gamma^T)}\sqrt{\gamma^{t'}(\gamma^{t'}+\gamma^{t'+1}-2\gamma^T)} &\leq 2\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\sqrt{\gamma^t(\gamma^t+\gamma^{t+1})}\sqrt{\gamma^{t'}(\gamma^{t'}+\gamma^{t'+1})} \\ &\leq 2\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\sqrt{\gamma^t(2\gamma^t)}\sqrt{\gamma^{t'}(2\gamma^{t'})} \\ &\leq 4\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\sqrt{\gamma^{2t}}\sqrt{\gamma^{2t'}} \\ &= 4\sum_{t=0}^{T-2}\gamma^t\sum_{t'=t+1}^{T-1}\gamma^{t'} \\ &\leq 4\left(\frac{1-\gamma^T}{1-\gamma}\right)^2\end{aligned}$$

Plugging everything together into Equation (52) leads to:

$$\frac{2\log(2/\delta)}{\Lambda}\left(\sum_{t=0}^{T-1}\sqrt{c_t}\right)^2 \leq \frac{2\log(2/\delta)}{\Lambda(1-\gamma)}\left(2\frac{1-\gamma^T}{1-\gamma} + 4\frac{(1-\gamma^T)^2}{(1-\gamma)^2}\right) \leq \frac{12\log(2/\delta)}{\Lambda(1-\gamma)^3}$$

Solving $\frac{12\log(2/\delta)}{\Lambda(1-\gamma)^3} \leq \epsilon^2$ for Λ leads to:

$$\Lambda = \mathcal{O}\left(\frac{\log(2/\delta)}{(1-\gamma)^3\epsilon^2}\right)$$

which concludes the first part of the proof.

Concerning the second part of the proof, we can bound:

$$\sum_{t=0}^{T-1}\gamma^t(\gamma^t + \gamma^{t+1} - 2\gamma^T) \leq 2\sum_{t=0}^{T-1}\gamma^{2t} \leq 2T$$

and:

$$\begin{aligned}
2\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\sqrt{\gamma^t(\gamma^t+\gamma^{t+1}-2\gamma^T)}\sqrt{\gamma^{t'}(\gamma^{t'}+\gamma^{t'+1}-2\gamma^T)} &\leq 4\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\gamma^t\gamma^{t'} \\
&\leq 4T\sum_{t=0}^{T-2}\gamma^t \\
&\leq 4T\frac{1-\gamma^T}{1-\gamma}
\end{aligned}$$

Plugging everything together into Equation (52) leads to:

$$\frac{2\log(2/\delta)}{\Lambda}\left(\sum_{t=0}^{T-1}\sqrt{c_t}\right)^2 \leq \frac{2\log(2/\delta)}{\Lambda(1-\gamma)}\left(2T + 4T\frac{1-\gamma^T}{1-\gamma}\right) \leq \frac{12T\log(2/\delta)}{\Lambda(1-\gamma)^2}$$

Solving $\frac{12T\log(2/\delta)}{\Lambda(1-\gamma)^2} \leq \epsilon^2$ for Λ leads to:

$$\Lambda = \mathcal{O}\left(\frac{T\log(2/\delta)}{(1-\gamma)^2\epsilon^2}\right)$$

which concludes the proof. □

B.2. Off-Policy Results

As for the on-policy setting, we begin by providing the unbiasedness results for Equation (9).

Theorem B.13. *Consider an optimization budget $\Lambda \geq T$ and a DCS such that $m_T \geq 1$. Consider policies $\pi_{\bar{\boldsymbol{\theta}}}, \pi_{\boldsymbol{\theta}} \in \Pi_\Theta$ such that $\pi_{\bar{\boldsymbol{\theta}}}(\cdot|s) \ll \pi_{\boldsymbol{\theta}}(\cdot|s)$ a.s. for every $s \in \mathcal{S}$, then:*

$$\mathop{\mathbb{E}}_{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}\left[\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})\right] = J(\bar{\boldsymbol{\theta}}). \tag{53}$$

Proof. Define $r_{t,\bar{\boldsymbol{\theta}}}$ as the expected t-th reward under policy $\pi_{\bar{\boldsymbol{\theta}}}$ and consider:

$$\begin{aligned}
\mathop{\mathbb{E}}_{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}\left[\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})\right] &= \mathop{\mathbb{E}}_{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}\left[\sum_{h=1}^{T}\sum_{i=1}^{m_h}\omega_{\bar{\boldsymbol{\theta}},\boldsymbol{\theta}}(\boldsymbol{\tau}_h^{(i)})\sum_{t=0}^{h-1}\gamma^t\frac{R(a_t^{(i)},s_t^{(i)})}{n_t}\right] \\
&= \sum_{t=1}^{T}m_h\mathop{\mathbb{E}}_{\boldsymbol{\tau}_h\sim p(\cdot|\boldsymbol{\theta},h)}\left[\omega_{\bar{\boldsymbol{\theta}},\boldsymbol{\theta}}(\boldsymbol{\tau}_h)\sum_{t=0}^{h-1}\gamma^t\frac{R(a_t,s_t)}{n_t}\right] \\
&= \sum_{t=1}^{T}m_h\mathop{\mathbb{E}}_{\boldsymbol{\tau}_h\sim p(\cdot|\bar{\boldsymbol{\theta}},h)}\left[\sum_{t=0}^{h-1}\gamma^t\frac{R(a_t,s_t)}{n_t}\right] \\
&= \sum_{h=1}^{T}m_h\sum_{t=0}^{h-1}\gamma^t\frac{r_{t,\bar{\boldsymbol{\theta}}}}{n_t},
\end{aligned}$$

where the first equality follows by the definition of $\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})$, the second from the linearity of the expectation together with the definition of the data generation process $p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta}, h)$, the third one from the IS property (Owen, 2013), and the forth one by the linearity of the expectation together with the definition of $r_{t,\bar{\boldsymbol{\theta}}}$. At this point, the rest of the proof follows directly from the one of Theorem B.1. □

We now continue by extending the high-probability confidence intervals of Metelli et al. (2018). In the rest of this section, we assume that rewards are bounded in $[-R_{\text{MAX}}, R_{\text{MAX}}]$ to allow for a direct comparison with Metelli et al. (2018). We also notice that, all the following results are derived based on the Cantelli's inequality, which is an appropriate choice for one-sided tail bounds. Two-sided tail bounds can be straightforwardly derived by using Chebyshev's inequality. For completeness, we begin by reporting the original result of Metelli et al. (2018).

Theorem B.14. *Let $\pi_{\bar{\boldsymbol{\theta}}}, \pi_{\boldsymbol{\theta}} \in \Pi_\Theta$ such that $\pi_{\bar{\boldsymbol{\theta}}}(\cdot|s) \ll \pi_{\boldsymbol{\theta}}(\cdot|s)$ a.s. for every $s \in \mathcal{S}$. Let us define the off-policy expected return estimator with K trajectories of horizon T collected with $\pi_{\boldsymbol{\theta}}$:*

$$\hat{J}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) = \frac{1}{K}\sum_{i=1}^{K} \omega_{\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}}(\boldsymbol{\tau}_T^{(i)}) \sum_{t=0}^{T-1} \gamma^t R(s_t^{(i)}, a_t^{(i)}) \tag{54}$$

Let $\Lambda = KT$, $\beta_\delta = \frac{1-\delta}{\delta}$ and $\phi = R_{MAX}\frac{1-\gamma^T}{1-\gamma}$, then, Then, with probability at least $1-\delta$ it holds that:

$$J(\bar{\boldsymbol{\theta}}) \geq \hat{J}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) - \phi\sqrt{\frac{T\beta_\delta d_2\left(p(\cdot|\bar{\boldsymbol{\theta}}, T)\|p(\cdot|\boldsymbol{\theta}, T)\right)}{\Lambda}}, \tag{55}$$

At this point, we are ready to provide our generalization of Theorem B.14 (which, for $R_{\text{MAX}} = 1$, reduces to Theorem 3.6 of Section 3).

Theorem B.15. *Consider $\pi_{\bar{\boldsymbol{\theta}}}$, $\pi_{\boldsymbol{\theta}} \in \Pi_\Theta$ such that $\pi_{\bar{\boldsymbol{\theta}}}(\cdot|s) \ll \pi_{\boldsymbol{\theta}}(\cdot|s)$ a.s. for all $s \in \mathcal{S}$. Consider an optimization budget $\Lambda \geq T$ and a generic DCS $\boldsymbol{m}$. Then, with probability at least $1-\delta$ it holds that:*

$$J(\bar{\boldsymbol{\theta}}) \geq \hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) - \sqrt{\beta_\delta \sum_{h=1}^{T} m_h \phi_h^2 d_2(p(\cdot|\bar{\boldsymbol{\theta}}, h)\|p(\cdot|\boldsymbol{\theta}, h))}, \tag{56}$$

where $\beta_\delta = \frac{1-\delta}{\delta}$ and $\phi_h \coloneqq R_{MAX}\sum_{t=0}^{h-1}\frac{\gamma^t}{n_t}$.

Proof. As in Metelli et al. (2018), we split the proof into two parts, i.e., first we upper bound the variance of the estimator, and then we make use of the Cantelli's inequality to prove Equation (56).

Let us start with the variance bound. Consider:

$$\underset{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}{\mathbb{V}\text{ar}}\left[\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})\right] = \underset{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}{\mathbb{V}\text{ar}}\left[\sum_{h=1}^{T}\sum_{i=1}^{m_h} \omega_{\bar{\boldsymbol{\theta}},\boldsymbol{\theta}}(\boldsymbol{\tau}_h^{(i)}) \sum_{t=0}^{h-1} \gamma^t \frac{R(a_t^{(i)}, s_t^{(i)})}{n_t}\right].$$

Since the different trajectories are independent, we can write:

$$\begin{aligned}
\underset{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}{\mathbb{V}\mathrm{ar}}\left[\sum_{h=1}^{T}\sum_{i=1}^{m_h}\omega_{\bar{\boldsymbol{\theta}},\boldsymbol{\theta}}(\boldsymbol{\tau}_h^{(i)})\sum_{t=0}^{h-1}\gamma^t\frac{R(a_t^{(i)},s_t^{(i)})}{n_t}\right] &= \sum_{h=1}^{T}m_h\underset{\boldsymbol{\tau}_h\sim p(\cdot|\boldsymbol{\theta},h)}{\mathbb{V}\mathrm{ar}}\left[\omega_{\bar{\boldsymbol{\theta}},\boldsymbol{\theta}}(\boldsymbol{\tau}_h)\sum_{t=0}^{h-1}\gamma^t\frac{R(s_t,a_t)}{n_t}\right] \\
&\leq \sum_{h=1}^{T}m_h\underset{\boldsymbol{\tau}_h\sim p(\cdot|\boldsymbol{\theta},h)}{\mathbb{E}}\left[\left(\omega_{\bar{\boldsymbol{\theta}},\boldsymbol{\theta}}(\boldsymbol{\tau}_h)\sum_{t=0}^{h-1}\gamma^t\frac{R(s_t,a_t)}{n_t}\right)^2\right] \\
&\leq \sum_{h=1}^{T}m_h\underset{\boldsymbol{\tau}_h\sim p(\cdot|\boldsymbol{\theta},h)}{\mathbb{E}}\left[\left(\omega_{\bar{\boldsymbol{\theta}},\boldsymbol{\theta}}(\boldsymbol{\tau}_h)\sum_{t=0}^{h-1}\gamma^t\frac{R_{\mathrm{MAX}}}{n_t}\right)^2\right] \\
&= \sum_{h=1}^{T}m_h\underset{\boldsymbol{\tau}_h\sim p(\cdot|\boldsymbol{\theta},h)}{\mathbb{E}}\left[\left(\omega_{\bar{\boldsymbol{\theta}},\boldsymbol{\theta}}(\boldsymbol{\tau}_h)\phi_h\right)^2\right] \\
&= \sum_{h=1}^{T}m_h\phi_h^2 d_2(p(\cdot|\bar{\boldsymbol{\theta}},h)||p(\cdot|\boldsymbol{\theta},h)),
\end{aligned}$$

where the last passage follows from the relationship between the moments of the importance weights and the Rényi divergence. This concludes the first part of the proof.

Concerning the second part, we start from the Cantelli's inequality applied on the random variable $\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})$, namely:

$$\mathbb{P}\left(\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) - J(\bar{\boldsymbol{\theta}}) \geq \alpha\right) \leq \frac{1}{1+\frac{\alpha^2}{\mathbb{V}\mathrm{ar}_{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}\left[\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})\right]}}$$

Set $\delta = \frac{1}{1+\frac{\alpha^2}{\mathbb{V}\mathrm{ar}_{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}\left[\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})\right]}}$ and consider the complementary event. Then, with probability at least $1-\delta$ it holds that:

$$\begin{aligned}
J(\bar{\boldsymbol{\theta}}) &\geq \hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) - \sqrt{\frac{1-\delta}{\delta}\underset{p_{\boldsymbol{m}}(\cdot|\boldsymbol{\theta})}{\mathbb{V}\mathrm{ar}}\left[\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})\right]} \\
&\geq \hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) - \sqrt{\beta_\delta\sum_{h=1}^{T}m_h\phi_h^2 d_2(p(\cdot|\bar{\boldsymbol{\theta}},h)\|p(\cdot|\boldsymbol{\theta},h))},
\end{aligned}$$

which concludes the proof.

□

As we can appreciate, using the uniform DCS in Equation (56) we recover exactly Equation (55) of Theorem B.14. At this point, one might be tempted to directly minimize Equation the confidence intervals around $J(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})$ as a function of $\boldsymbol{m}$ to obtain a tighter high-probability bound. However, as noted in Metelli et al. (2018), computing the Rényi divergence over the trajectory space requires both the approximation of a complex integral, and, for stochastic environments, the knowledge of the transition kernel P of the underlying MDP. Therefore, to derive a tractable expression that can be optimized as a function of the DCS, we further bound each term $d_2(p(\cdot|\bar{\boldsymbol{\theta}},h)\|p(\cdot|\boldsymbol{\theta},h))$ with $d_2(p(\cdot|\bar{\boldsymbol{\theta}},T)\|p(\cdot|\boldsymbol{\theta},T))$, which is justified by the following result.

Lemma B.16. *Consider two policies $\pi_{\bar{\boldsymbol{\theta}}},\pi_{\boldsymbol{\theta}}\in\Pi_\Theta$ such that $\pi_{\bar{\boldsymbol{\theta}}}\ll\pi_{\boldsymbol{\theta}}$ a.s. for every $s\in\mathcal{S}$. Consider $h\in\{1,\dots,T-2\}$, then:*

$$d_2(p(\cdot|\bar{\boldsymbol{\theta}},h)\|p(\cdot|\boldsymbol{\theta},h)) \leq d_2(p(\cdot|\bar{\boldsymbol{\theta}},h+1)\|p(\cdot|\boldsymbol{\theta},h+1)).$$

Proof. Focus on $h+1$. Due to the link between $d_2(p(\cdot|\bar{\boldsymbol{\theta}},h+1)\|p(\cdot|\boldsymbol{\theta},h+1))$ and the second moment of the importance weights, we have that:

$$d_2(p(\cdot|\bar{\boldsymbol{\theta}}, h+1)\|p(\cdot|\boldsymbol{\theta}, h+1)) = \mathop{\mathbb{E}}_{\boldsymbol{\tau}_{h+1}\sim p(\cdot|\boldsymbol{\theta},h+1)}\left[\prod_{t=0}^{h}\left(\frac{\pi_{\bar{\boldsymbol{\theta}}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}}(a_t|s_t)}\right)^2\right]$$
$$= \mathop{\mathbb{E}}_{\boldsymbol{\tau}_{h+1}\sim p(\cdot|\boldsymbol{\theta},h+1)}\left[\prod_{t=0}^{h-1}\left(\frac{\pi_{\bar{\boldsymbol{\theta}}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}}(a_t|s_t)}\right)^2\left(\frac{\pi_{\bar{\boldsymbol{\theta}}}(a_h|s_h)}{\pi_{\boldsymbol{\theta}}(a_h|s_h)}\right)^2\right].$$

Now, since $\boldsymbol{\tau}_{h+1} = (s_0, a_0, \ldots, s_h, a_h, s_{h+1})$, we can write the last expectation as:

$$\mathop{\mathbb{E}}_{(s_0,a_0,\ldots,s_{h-1},a_{h-1})\sim p(\cdot|\boldsymbol{\theta},h)}\left[\prod_{t=0}^{h-1}\left(\frac{\pi_{\bar{\boldsymbol{\theta}}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}}(a_t|s_t)}\right)^2 \mathop{\mathbb{E}}_{(s_h,a_h,s_{h+1})\sim p_h(\cdot|\boldsymbol{\theta},h+1)}\left[\left(\frac{\pi_{\bar{\boldsymbol{\theta}}}(a_h|s_h)}{\pi_{\boldsymbol{\theta}}(a_h|s_h)}\right)^2\right]\right]. \tag{57}$$

where with $p_h(\cdot|\boldsymbol{\theta}, h+1)$ we denote the h-th step (i.e., the last one) in a trajectory of length $h+1$. With a little abuse of notation, we drop the dependency on $p_h(\cdot|\boldsymbol{\theta}, h+1)$ and we write:

$$\mathop{\mathbb{E}}_{s_h,a_h}\left[\left(\frac{\pi_{\bar{\boldsymbol{\theta}}}(a_h|s_h)}{\pi_{\boldsymbol{\theta}}(a_h|s_h)}\right)^2\right] = \mathop{\mathbb{E}}_{s_h}\left[\mathop{\mathbb{E}}_{a_h\sim\pi_{\boldsymbol{\theta}}(\cdot|s_h)}\left[\left(\frac{\pi_{\bar{\boldsymbol{\theta}}}(a_h|s_h)}{\pi_{\boldsymbol{\theta}}(a_h|s_h)}\right)^2\right]\right] \geq \inf_{s\in\mathcal{S}}\mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\left(\frac{\pi_{\bar{\boldsymbol{\theta}}}(a|s)}{\pi_{\boldsymbol{\theta}}(a|s)}\right)^2\right] \geq 1, \tag{58}$$

where the last inequality follows from the fact that $\mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\left(\frac{\pi_{\bar{\boldsymbol{\theta}}}(a|s)}{\pi_{\boldsymbol{\theta}}(a|s)}\right)^2\right]$ can be interpreted as the exponentiated Rényi divergence with $\alpha = 2$ at state s. At this point, plugging Equation (58) in Equation (57), it follows that:

$$d_2(p(\cdot|\bar{\boldsymbol{\theta}}, h+1)\|p(\cdot|\boldsymbol{\theta}, h+1)) \geq \mathop{\mathbb{E}}_{(s_0,a_0,\ldots,s_{h-1},a_{h-1})\sim p(\cdot|\boldsymbol{\theta},h)}\left[\prod_{t=0}^{h-1}\left(\frac{\pi_{\bar{\boldsymbol{\theta}}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}}(a_t|s_t)}\right)^2\right]$$
$$= \mathop{\mathbb{E}}_{\boldsymbol{\tau_h}\sim p(\cdot|\boldsymbol{\theta},h)}\left[\prod_{t=0}^{h-1}\left(\frac{\pi_{\bar{\boldsymbol{\theta}}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}}(a_t|s_t)}\right)^2\right]$$
$$= d_2(p(\cdot|\bar{\boldsymbol{\theta}}, h)\|p(\cdot|\boldsymbol{\theta}, h)),$$

which concludes the proof. □

At this point, combining Lemma B.2 with Lemma B.16 and Theorem B.15, we obtain the following contrained optimization problem.

$$\begin{aligned}\min_{\boldsymbol{n}} \quad & \sqrt{\beta_\delta d_2(p(\cdot|\bar{\boldsymbol{\theta}}, T)||p(\cdot|\boldsymbol{\theta}, T))\sum_{t=0}^{T-1}\frac{c_t}{n_t}} \\ \text{s.t.} \quad & n_t \geq n_{t+1}, \quad \forall t \in \{0,\ldots,T-2\} \\ & \sum_{t=0}^{T-1} n_t = \Lambda \\ & n_t \in \mathbb{N}_+, \quad \forall t \in \{0,\ldots,T-1\}\end{aligned} \tag{59}$$

Notice that (59) is equivalent, up to constant factors, to the one of the on-policy case. Consequently, it can be solved using the same methodology applied in the previous section. More specifically, following the same proof scheme, it is possible to derive the following result.

Theorem B.17. *Consider an optimization budget $\Lambda > T$, let $\tilde{\boldsymbol{n}}^*$ be the approximately optimal DCS given in Definition B.8, and let $\boldsymbol{n}^*$ be the optimal solution of the integer optimization problem* (59)*. Moreover, let $f(\boldsymbol{n}) = \sqrt{\beta_\delta d_2(p(\cdot|\bar{\boldsymbol{\theta}}, T)||p(\cdot|\boldsymbol{\theta}, T))\sum_{t=0}^{T-1}\frac{c_t}{n_t}}$. Then,*

$$f(\boldsymbol{n}^*) \leq f(\tilde{\boldsymbol{n}}^*) \leq \sqrt{2} f(\boldsymbol{n}^*) \tag{60}$$

B.3. Further analysis

In Theorem B.10, we have seen that $f(\tilde{\boldsymbol{n}}^*) \leq \sqrt{2} f(\boldsymbol{n}^*)$. We have now ask ourselves if we can obtain tighter values for the constant. The following results provides a positive answer.

Proposition B.18. *Consider an optimization budget $\Lambda > T$, let $\tilde{\boldsymbol{n}}^*$ be the approximately optimal DCS given in Defintion B.8, let $\boldsymbol{n}^*$ be the optimal solution of the integer optimization problem* (6)*, and let $\bar{\boldsymbol{n}}^*$ be the solution of the convex relaxation of* (6) *given in Lemma B.7. Define $\mathcal{X} = \{x \in (0,1) : \bar{n}^*_{T-1} \geq \frac{1}{1-x}\}$. Then, if $\bar{n}^*_{T-1} > 1$ holds, we have that:*

$$f(\tilde{\boldsymbol{n}}^*) \leq \min_{x \in \mathcal{X}} \sqrt{\frac{1}{x}} f(\boldsymbol{n}^*) \tag{61}$$

Proof. Let us analyze: $\frac{f(\tilde{\boldsymbol{n}}^*)}{f(\boldsymbol{n}^*)}$. For the same reasoning of Theorem B.10, we have that:

$$\frac{f(\tilde{\boldsymbol{n}}^*)}{f(\boldsymbol{n}^*)} \leq \frac{f(\tilde{\boldsymbol{n}}^*)}{f(\bar{\boldsymbol{n}}^*)} \tag{62}$$

Then, we provide an upper bound on $f(\tilde{\boldsymbol{n}}^*)$ that holds whenever $\bar{n}^*_{T-1} \geq 1$. More specifically, consider a generic $x \in \mathcal{X}$:[15]

$$f(\tilde{\boldsymbol{n}}^*) = \sqrt{\frac{1}{2}\log(2/\delta) \sum_{t=0}^{T-1} \frac{c_t}{\tilde{n}^*_t}} \leq \sqrt{\frac{1}{2}\log(2/\delta) \sum_{t=0}^{T-1} \frac{c_t}{\bar{n}^*_t - 1}} \leq \sqrt{\frac{1}{2}\log(2/\delta) \sum_{t=0}^{T-1} \frac{c_t}{x\bar{n}^*_t}} = \sqrt{\frac{1}{x}} f(\bar{\boldsymbol{n}}^*) \tag{63}$$

where, in the first inequality we have used Lemma B.9 together with $\bar{n}^*_{T-1} > 1$ [16], and in the second one we have used the definition of $\mathcal{X}$.

Plugging Equation (63) into (62) concludes the proof. □

Proposition B.18 provides a tighter upper bound that depends on the number of samples allocated to $\bar{n}^*_{T-1}$ by the solution of the convex relaxation of (6). Due to Lemma B.7, we know that there is a tight relationship between $\bar{n}^*_t$ and the available budget Λ. More specifically, we can appreciate that, as the budget increase, so does $\bar{n}^*_{T-1}$. Due to Equation (61), this, in turn, implies tighter upper bounds on the quality of the approximately optimal DCS. As an example, suppose that $\bar{n}^*_{T-1} = 100$. Then, we have that:

$$f(\tilde{\boldsymbol{n}}^*) \leq \sqrt{\frac{100}{99}} f(\boldsymbol{n}^*)$$

We now provide some variance analysis for settings in which rewards are gathered at the end of the episode.

Proposition B.19. *Consider the MDP of Figure 3 together with the indicated policy. Fix a DCS $\boldsymbol{m}$ such that $m_T \geq 1$. Then, it holds that:*

$$\underset{p_{\boldsymbol{m}}(\cdot|\bar{\boldsymbol{\theta}})}{\mathbb{V}\mathrm{ar}} \left[\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\bar{\boldsymbol{\theta}})\right] = \frac{\gamma^{2(T-1)}}{n_{T-1}}. \tag{64}$$

Proof.

$$\underset{p_{\boldsymbol{m}}(\cdot|\bar{\boldsymbol{\theta}})}{\mathbb{V}\mathrm{ar}} \left[\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\bar{\boldsymbol{\theta}})\right] = \underset{p_{\boldsymbol{m}}(\cdot|\bar{\boldsymbol{\theta}})}{\mathbb{V}\mathrm{ar}} \left[\sum_{h=1}^{T} \sum_{i=1}^{m_h} \sum_{t=0}^{h-1} \gamma^t \frac{R(s_t^{(i)}, a_t^{(i)})}{n_t}\right]$$

[15] Notice that when $\bar{n}^*_{T-1} > 1$, $\mathcal{X}$ is a non-empty set.

[16] Notice that $\bar{n}^*_{T-1} > 1$ implies $\bar{n}^*_t > 1$ for all $t \in \{0, \ldots, T-1\}$.

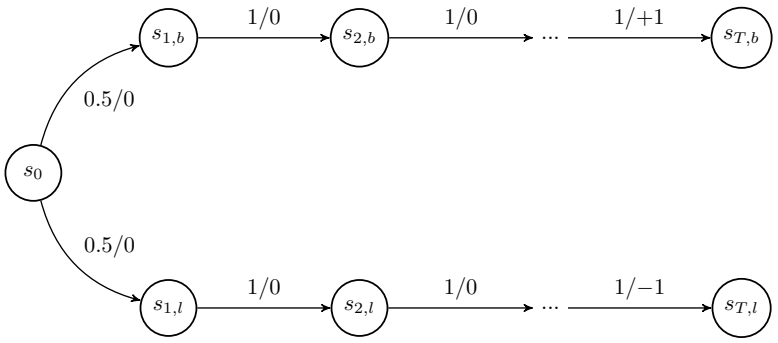

Figure 3. MDP example with a fixed policy $\pi_{\bar{\boldsymbol{\theta}}}$ in which rewards are gathered only at the end of the episode. Each edge reports the probability of taking that action, together with its associated reward. For states in which there is a single edge that is followed with probability 1 (e.g., $s_{1,l}$), other actions have been masked.

Since the different trajectories are independent, we can write:

$$\begin{aligned}
\underset{p_{\boldsymbol{m}}(\cdot|\bar{\boldsymbol{\theta}})}{\mathbb{V}\mathrm{ar}}\left[\sum_{h=1}^{T}\sum_{i=1}^{m_h}\sum_{t=0}^{h-1}\gamma^t\frac{R(s_t^{(i)},a_t^{(i)})}{n_t}\right] &= \sum_{h=1}^{T} m_h \underset{p(\cdot|\bar{\boldsymbol{\theta}},h)}{\mathbb{V}\mathrm{ar}}\left[\sum_{t=0}^{h-1}\gamma^t\frac{R(s_t,a_t)}{n_t}\right] \\
&= \sum_{h=1}^{T} m_h \underset{p(\cdot|\bar{\boldsymbol{\theta}},h)}{\mathbb{E}}\left[\left(\sum_{t=0}^{h-1}\gamma^t\frac{R(s_t,a_t)}{n_t} - \underset{p(\cdot|\bar{\boldsymbol{\theta}},h)}{\mathbb{E}}\left[\sum_{t=0}^{h-1}\gamma^t\frac{R(s_t,a_t)}{n_t}\right]\right)^2\right] \\
&= \sum_{h=1}^{T} m_h \underset{p(\cdot|\bar{\boldsymbol{\theta}},h)}{\mathbb{E}}\left[\left(\sum_{t=0}^{h-1}\gamma^t\frac{R(s_t,a_t)}{n_t}\right)^2\right] \\
&= \sum_{h=1}^{T} m_h \underset{p(\cdot|\bar{\boldsymbol{\theta}},h)}{\mathbb{E}}\left[\sum_{t=0}^{h-1}\gamma^{2t}\frac{R(s_t,a_t)^2}{n_t^2} + \sum_{t=0}^{h-2}\sum_{t'=t+1}^{h-1}\gamma^{t+t'}\frac{R(s_t,a_t)R(s_{t'},a_{t'})}{n_t n_{t'}}\right] \\
&= m_T\frac{\gamma^{2(T-1)}}{n_{T-1}^2} \\
&= \frac{\gamma^{2(T-1)}}{n_{T-1}}
\end{aligned}$$

which concludes the proof. □

From Equation (64), we can make the following consideration. As noticed in Section 3, the uniform strategy is intuitively a good choice for settings such as the one of Figure 3. Indeed, the only relevant data is gathered at the end of the episode, and, therefore, we need to interact with the environment as much as possible at time $t = T - 1$. The variance of the uniform strategy will be given by:

$$\frac{\gamma^{2(T-1)}T}{\Lambda} \tag{65}$$

At this point, what can we say about our approximately optimal DCS? The following proposition summarizes the result.

Proposition B.20. *Consider the MDP of Figure 3 together with the indicated policy. Suppose that* $T \geq \frac{\log\left(\frac{1}{(1-\gamma)\gamma}\right)}{\log\left(\frac{1}{\gamma}\right)}$ *holds. Consider* $\tilde{\boldsymbol{m}}^*$ *as in Definition* (B.8)*. Then, for sufficiently large values of* Λ*, it holds that:*

$$\underset{p_{\tilde{\boldsymbol{m}}^*}(\cdot|\bar{\boldsymbol{\theta}})}{\mathbb{V}\mathrm{ar}}\left[\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\bar{\boldsymbol{\theta}})\right] \leq \frac{2\gamma^{\frac{1}{2}(T-1)}T}{\Lambda}. \tag{66}$$

Proof. Due to Proposition B.19, we are interested in studying:

$$\frac{\gamma^{2(T-1)}}{\tilde{n}^*_{T-1}}$$

Let $\bar{\boldsymbol{n}}^*$ be the solution of the convex relaxation given in Lemma B.7. Then,

$$\frac{\gamma^{2(T-1)}}{\bar{n}^*_{T-1}} \geq \frac{\gamma^{2(T-1)}}{\tilde{n}^*_{T-1}+1} \geq \frac{\gamma^{2(T-1)}}{2\tilde{n}^*_{T-1}}$$

Which leads to:

$$\frac{\gamma^{2(T-1)}}{\tilde{n}^*_{T-1}} \leq \frac{2\gamma^{2(T-1)}}{\bar{n}^*_{T-1}} \tag{67}$$

At this point, due to Lemma B.7, for sufficiently large values of Λ, we can substitute $\bar{n}^*_{T-1}$ with:

$$\frac{2\gamma^{2(T-1)}}{\bar{n}^*_{T-1}} = \frac{2\gamma^{2(T-1)}\sum_{t=0}^{T-1}\sqrt{c_t}}{\Lambda\sqrt{c_{T-1}}} \leq \frac{2\gamma^{2(T-1)}T}{\Lambda\sqrt{\gamma^{T-1}(\gamma^{T-1}-\gamma^T)}} = \frac{2\gamma^{2(T-1)}T}{\Lambda\sqrt{\gamma^{2(T-1)}(1-\gamma)}} \tag{68}$$

where in the inequality step we have used the definition of c_t. Moreover, since $T \geq \frac{\log\left(\frac{1}{(1-\gamma)\gamma}\right)}{\log\left(\frac{1}{\gamma}\right)}$ holds by assumption, we have that $(1-\gamma) \geq \gamma^{T-1}$, therefore, we can further upper-bound Equation (68) with:

$$\frac{2\gamma^{2(T-1)}T}{\Lambda\sqrt{\gamma^{3(T-1)}}} = \frac{2\gamma^{\frac{1}{2}(T-1)}T}{\Lambda}$$

which concludes the proof. □

We now make some remarks both on the setting and the comparison between the uniform strategy and our approximately optimal DCS $\tilde{\boldsymbol{m}}^*$. First of all, from Equation (64) we notice that the variance tends to 0 with an exponential rate w.r.t. the horizon T, meaning that, when the horizon is sufficiently large, any method will enjoy numerically low variance. Furthermore, Equation (66) shows that, for sufficiently large values of Λ, the variance of $\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\bar{\boldsymbol{\theta}})$ displays a very similar behavior w.r.t. the one obtained by the uniform approach; the only difference, indeed, stands in a different power of γ. Furthermore, we remark that, in such sparse reward settings, γ is usually very close to 1 to avoid nullifying the rewards that are gathered at the end of the episode. However, whenever this happens, $\tilde{\boldsymbol{m}}^*$ will tend to the uniform strategy.

Finally, we conclude with some remarks on off-policy PAC bounds that can be derived using our approximately optimal DCS. More specifically, given some desired accuracy level $\epsilon > 0$, we aim at answering the following question: what is the minimum amount of budget Λ such that $|\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) - J(\bar{\boldsymbol{\theta}})| \leq \epsilon$ holds with probability at least $1-\delta$? First of all, to answer this question we notice that we need to rely on two-sided tail bounds. To this end, it is sufficient to modify the proof of Theorem 3.6 by using Chebyshev's inequality, rather than Cantelli's one. The result that it possible to obtain is the following one:

$$|J(\bar{\boldsymbol{\theta}}) - \hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})| - \sqrt{\bar{\beta}_\delta \sum_{h=1}^{T} m_h \phi_h^2 d_2(p(\cdot|\bar{\boldsymbol{\theta}},h)\|p(\cdot|\boldsymbol{\theta},h))}, \tag{69}$$

where $\bar{\beta}_\delta = \frac{1}{\delta}$ and $\phi_h \coloneqq \sum_{t=0}^{h-1}\frac{\gamma^t}{n_t}$. At this point, define, for brevity: $\omega = d_2(p(\cdot|\bar{\boldsymbol{\theta}},T)\|p(\cdot|\boldsymbol{\theta},T))$. It is easy to see, that the uniform-in-the-horizon $\bar{\boldsymbol{m}}$ strategy requires:

$$\Lambda = \mathcal{O}\left(\frac{\omega\bar{\beta}_\delta T}{(1-\gamma)^2\epsilon^2}\right), \tag{70}$$

to satisfy $|\hat{J}_{\bar{\boldsymbol{m}}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) - J(\bar{\boldsymbol{\theta}})| \leq \epsilon$ with high probability. Concerning our approximately optimal DCS, instead, under the assumption that $\omega\bar{\beta}_\delta c_0 \geq 4T\epsilon^2$ holds, we have that:

$$\Lambda = \mathcal{O}\left(\min\left\{\frac{\omega\bar{\beta}_\delta}{(1-\gamma)^3\epsilon^2}, \frac{\omega\bar{\beta}_\delta T}{(1-\gamma)^2\epsilon^2}\right\}\right). \tag{71}$$

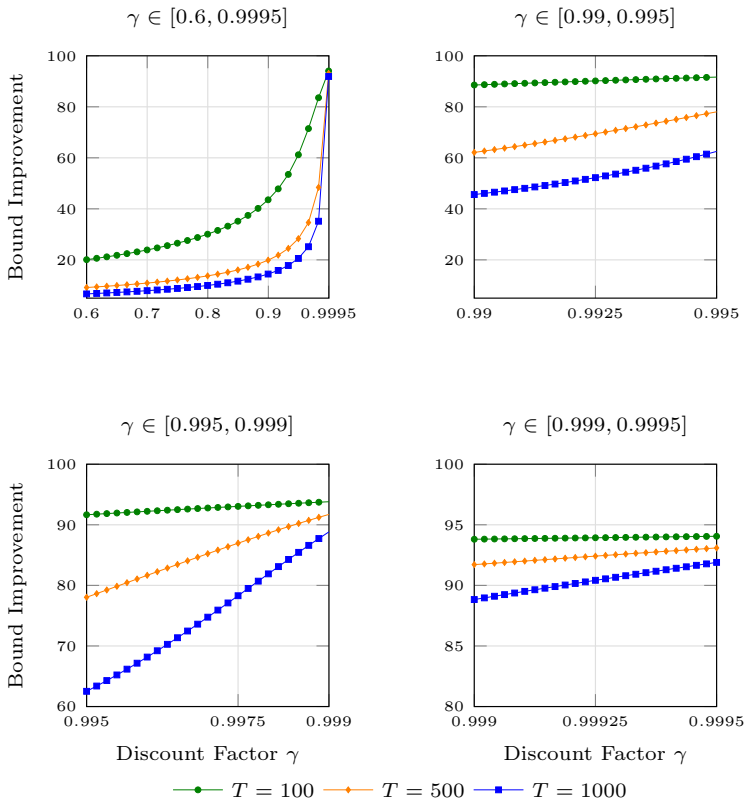

Figure 4. Visualization of the confidence interval improvements reported as a function of γ for different values of T, when using $\Lambda = 10k$.

The proof follows directly from the ones of Theorem 3.4 by substituting $\frac{1}{2}\log\left(\frac{1}{\delta}\right)$ with $\bar{\beta}_\delta\omega$. The main conceptual difference between these bounds and the on-policy ones is that these results are only *descriptive* and not *prescriptive* (i.e., we cannot make a budget decision accordingly). Indeed, the Rényi divergence depends on the unknown transition model, which is unknown to the learner, and the two considered policies. To get a prescriptive result, one might look for an upper bound on this divergence. However, we notice that whatever upper-bound one might use, our method will still enjoy robustness and improvements over the uniform strategy. We thank the reviewer for highlighting this interesting detail. We will include such comments in the final version of the paper.

C. Visualizations

In this Section, we provide some visualizations of the improvement in the confidence intervals in Figure 4 when varying γ and T, and keeping fixed Λ. The improvement is reported as $\frac{100 f(\tilde{\boldsymbol{n}}^*)}{f(\boldsymbol{n}_u)}$, where $\boldsymbol{n}_u$ is the uniform allocation strategy and $f(\boldsymbol{n}) = \sqrt{\frac{1}{2}\log(2/\delta)\sum_{t=0}^{T-1}\frac{c_t}{n_t}}$. Moreover, Figure 5 provides visualizations on $\tilde{\boldsymbol{n}}^*$ as a function of γ and for different values of Λ when $T = 10$.

D. Additional details on POIS and TT-POIS

D.1. Pseudo-code and other details

The pseudo-code for our algorithm, TT-POIS, can be found in Algorithm 1. As one can notice, by replacing $\tilde{\boldsymbol{m}}^*$ with the uniform-in-the-horizon DCS, we recover the original pseudo-code of POIS (Metelli et al., 2018). We remark that, as in POIS, δ is treated as an hyper-parameter, and that in Line 6, the step size is computed online via line search.

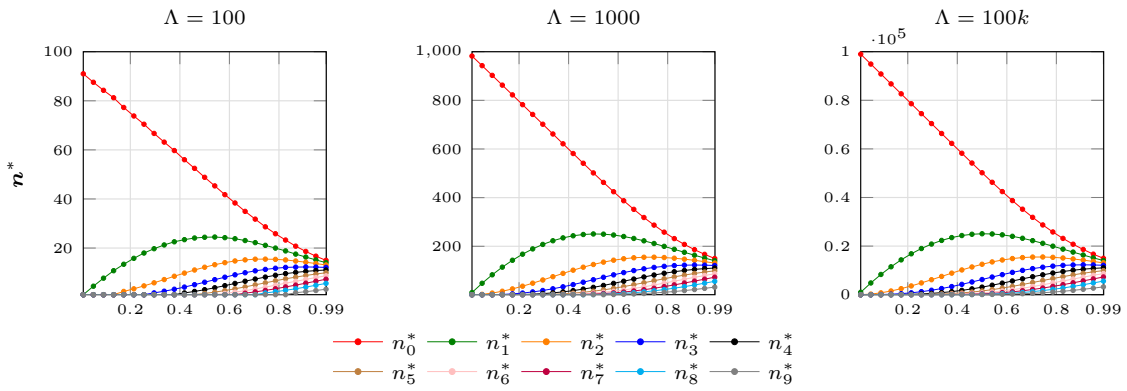

Figure 5. Visualization of $\boldsymbol{n}^*$ for different values of γ and Λ when using $T = 10$.

Algorithm 1 Truncating Trajectories in Policy Optimization via Importance Sampling (TT-POIS)

Require: Optimization budget Λ, confidence level δ
1: Initialize $\boldsymbol{\theta}_0^0$ arbitrarily
2: Compute $\tilde{\boldsymbol{m}}^*$ as in Definition B.8
3: **for** $j = 0, 1, 2, \dots$ **do**
4: Collect dataset $\mathcal{D} \sim p_{\tilde{\boldsymbol{m}}^*}(\cdot|\boldsymbol{\theta}_0^j)$
5: **for** $k = 0, 1, 2, \dots$ **do**
6: Compute $\nabla\mathcal{L}_\delta(\boldsymbol{\theta}_k^j/\boldsymbol{\theta}_0^j)$ and α_k
7: $\boldsymbol{\theta}_{k+1}^j = \boldsymbol{\theta}_k^j + \alpha_k \nabla\mathcal{L}(\boldsymbol{\theta}_k^j/\boldsymbol{\theta}_0^j)$
8: **end for**
9: **end for**

We now recall the definition of the objective function $\mathcal{L}_\delta$ provided in Section 4, namely:

$$\mathcal{L}_\delta(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) \coloneqq \hat{J}_{\tilde{\boldsymbol{m}}^*}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta}) - \sqrt{\beta_\delta \sum_{h=1}^{T} \tilde{m}_h^*(\tilde{\phi}_h^*)^2 \hat{d}_2(p(\cdot|\bar{\boldsymbol{\theta}}, h)|p(\cdot|\boldsymbol{\theta}, h))}, \tag{72}$$

where $\tilde{\phi}_h^* = \sum_{t=0}^{h-1} \frac{\gamma^t}{\tilde{n}_t^*}$ and $\hat{d}_2(p(\cdot|\bar{\boldsymbol{\theta}}, h)|p(\cdot|\boldsymbol{\theta}, h))$ is a sampled-based approximation for $d_2(p(\cdot|\bar{\boldsymbol{\theta}}, h)|p(\cdot|\boldsymbol{\theta}, h))$. More specifically,

$$\hat{d}_2(p(\cdot|\bar{\boldsymbol{\theta}}, h)|p(\cdot|\boldsymbol{\theta}, h)) = \frac{1}{n_h} \sum_{i=h}^{T} \sum_{j=1}^{\tilde{m}_i^*} \prod_{t=0}^{h-1} d_2(\pi_{\bar{\boldsymbol{\theta}}}(\cdot|s_{\tau_i^{(j)},t}) \| \pi_{\boldsymbol{\theta}}(\cdot|s_{\tau_i^{(j)},t})) \tag{73}$$

Notice that Equation (73), when applied with the uniform DCS, recovers the same approximation used in Metelli et al. (2018) (see their Equation 41).

As an additional comment, we notice that in Section 3, Equation (72) has been presented for rewards in $[0, 1]$. For the more general case in which rewards are defined in $[-R_{\text{MAX}}, R_{\text{MAX}}]$, it is sufficient to replace $\tilde{\phi}_h^* = \sum_{t=0}^{h-1} \frac{\gamma^t}{\tilde{n}_t^*}$ with $\tilde{\phi}_h^* = R_{\text{MAX}} \sum_{t=0}^{h-1} \frac{\gamma^t}{\tilde{n}_t^*}$ (see Theorem B.15).

D.2. Implementation Details

Our implementation follows directly from the original one of POIS (Metelli et al., 2018). More specifically, the line search method adopted for performing the update (Line 6 in Algorithm 1) is the same of Metelli et al. (2018). In this sense, the reader can refer to Appendix E.1 of Metelli et al. (2018) for futher details. Compared to Metelli et al. (2018), however, we introduce the two following hyper-parameters that have been used, in our experiments, both for POIS and TT-POIS.

Minimum-maximum empirical reward In the original version of POIS (Metelli et al., 2018) (and also in our experiments), in $\mathcal{L}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})$, R_{MAX} is replaced with the maximum empirical reward $\hat{R}_{\text{MAX}}$ that is collected at the current training iteration. This leads to a further adaptivity of $\mathcal{L}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})$. However, in domain such as the Reacher, where the rewards tend to be close to 0 when good policies are learnt, using the maximum empirical reward might lead to numerical instabilities. Indeed, in these situations, the adaptive trust region will approach 0, and both POIS and TT-POIS will simply maximimize $\hat{J}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})$, with no control on the variance of the importance weights. For this reason, we define an additional hyper-parameter, $R_{\text{MIN-MAX}}$, that defines a minimum threshold for $\hat{R}_{\text{MAX}}$. If $\hat{R}_{\text{MAX}}$ falls below $R_{\text{MIN-MAX}}$, then $R_{\text{MIN-MAX}}$ will be used in $\mathcal{L}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})$.

Importance weights clipping When employing POIS and TT-POIS in domains with discrete actions (e.g., the supply chain), it might happen that in some states some actions are highly sub-optimal. In this case, even if we are controlling the variance of the importance weights, the objective function might lead to shrink their probability to 0. In training, this can results in NAN gradients and numerical instabilitities. For this reason, we clip the importance weights in $\hat{J}_{\boldsymbol{m}}(\bar{\boldsymbol{\theta}}/\boldsymbol{\theta})$ with an hyper-parameter IW_c.

E. Experiment Details and Additional Results

In this Section, we provide further details on the experiments and additional results. More specifically:

- Section E.1 provides an in-depth description for each environment that has been considered.
- Section E.2 provides results that purely focus on the evaluation setting of Section 3.
- Section E.3 provides ablation experiments on the policy optimization setting.
- Section E.4 provides additional results on the experimental setting of Section 5. More specifically, results with additional values of Λ and γ are presented.
- Section E.5 provides additional results on the experimental setting of Section 5. More specifically, the undiscounted return metric is reported.
- Section E.6 provides additional results on the Reacher domain where T and γ varies jointly while keeping γ^T roughly constant.
- Section E.7 reports hyper-parameters and other practical details.

E.1. Environment Details

E.1.1. Evaluation Domain

We now provide a description of the environment that is used to conduct evaluation experiments, whose results are presented in Section E.2. More specifically, we designed a domain with the following features:

- The performance of any policy can be easily computed in closed form.
- It can easily generalize to any value of T so that we can study the behavior of the algorithm varying T.

Given these general features, we designed the following environment. The state is described solely by the integer variable t, which represents the step in which the action is taken. The action space is discrete, with 2 possible actions. Concerning the reward function, since we want it to generalize to any horizon T, we made the following design choices. We restricted ourselves to $T \in \{100, 1000, 2000\}$. Then, focus for the sake of exposition on $T = 100$. Define $\boldsymbol{g}_1 = (1, 4, 3, 1, 1.5, 0.4, 4, 4.1, 3, 2, 4)$ and $\boldsymbol{g}_2 = (4, 1, 1, 3, 4, 1.5, 0.1, 5, 1, 1, 4)$. Then, if $t \notin \{0, 10, 20, \dots, 90, 99\}$, $R(a_1, s_t) = 0$ and $R(a_2, s_t) = 0$. If $t \in \{0, 10, 20, \dots, 90, 99\}$, denote with $i(t)$ the corresponding index of the element t within the vector $\{0, 10, 20, \dots, 90, 99\}$; then $R(a_1, s_t) = \mathcal{N}\left(g_{1,i(t)}, 0.1\right)$ and $R(a_2, s_t) = \mathcal{N}\left(g_{2,i(t)}, 0.1\right)$.

Similar reasoning extends to the cases in which T is equal to 1000 and 2000 by considering the vectors $\{0, 100, 200, 300, 400, 500, 600, 700, 800, 900, 999\}$ and $\{0, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 1999\}$ respectively. Further details can be found in the code base we provide.

E.1.2. CORRIDOR DOMAIN

In order to compare POIS and TT-POIS on very similar domains but with different reward functions we design the following experiment. More specifically, the domain represents a corridor: the agent starts in the middle, and needs to reach the right extreme. To this end, it has two possible actions: "go left" and "go right", which both succeed with a certain probability. Then, we consider the following reward functions:

- Sparse reward: the reward is equal to 1 only if the agent has reached the right extreme of the corridor, and 0 otherwise.
- Dense reward: if the selected action is "go right", the agent receives with high probability reward 1, while if the selected action is "go left", it receives with high probability reward -1.

More formally, we consider a continuous state space $\mathcal{S} \in [-x_{\text{MAX}}, x_{\text{MAX}}]$ for some $x_{\text{MAX}} > 0$. The initial state is fixed, and equal to 0, namely $s_0 = 0$. Then, denote with a_1 the action "go right" and with a_2 the action "go left". Then, consider $\bar{p} \in (0.5, 1)$ (i.e., the probability of success of a given action). Let x_t be the current state, then, if $a_t = a_1$, $x_{t+1} = 1 + x_t + q$ where $q \sim \mathcal{N}(0, 0.1)$ with probability $\bar{p}$, and $x_{t+1} = -1 + x_t + q$ where $q \sim \mathcal{N}(0, 0.1)$ with probability $1 - \bar{p}$. Similarly, if $a_t = a_2$, $x_{t+1} = -1 + x_t + q$ where $q \sim \mathcal{N}(0, 0.1)$ with probability $\bar{p}$, and $x_{t+1} = 1 + x_t + q$ where $q \sim \mathcal{N}(0, 0.1)$ with probability $1 - \bar{p}$. Then, we also clip the value of x_t to be in the specified range; namely we clip x_t in $[-x_{\text{MAX}}, x_{\text{MAX}}]$. Then, let the goal state be $x_g = x_{\text{MAX}}$. We say that the goal is reached whenever $|x - x_g| < 0.5$ holds. Furthermore, states for which the goal is reached are modelled as absorbing states.

For what concerns the reward function, instead, let us first consider the sparse reward setting. In this case, let the goal state be $x_g = x_{\text{MAX}}$. Then, $R(x_t, a_t) = 1$ if $|x_t - x_g| < 0.5$, 0 otherwise. Furthermore, we set $T = 100$ and $x_g = 12$.[17]

For the dense reward setting, instead, $R(x, \cdot) = 0$ for x such that $|x - x_g| < 0.5$ holds. For x such that $|x - x_g| < 0.5$ does not hold instead, $R(\cdot, a_1)$ is 0.2 with probability $\bar{p}$ and -0.2 with probability $1 - \bar{p}$. Moreover, $R(\cdot, a_2)$ is -0.2 with probability $\bar{p}$ and 0.2 with probability $1 - \bar{p}$. Furthermore, we set $T = 1000$ and $x_g = 1000$.

Further details can be found in the code base we provide.

E.1.3. DAM CONTROL

We now provide additional details on the dam environment. First, we detail the adopted parameters of the environment, then we describe the state observed by the agents and the action space considered.

The parameters of the environment are the default ones of (Tirinzoni et al., 2018):

- The demand D is fixed and equal to 10.
- The inflow profile i_t is a period function (i.e., Figure 6) plus Gaussian noise with $\sigma = 2$.
- The initial storage s_0 is set to 200.
- The flooding threshold F is set to 300.
- The reward is computed as $-c_1 \max\{0, s_t - F\} - c_2 \max\{0, D - a_t\}^2$ with $c_1 = c_2 = 0.5$. Moreover, rewards are rescaled with 0.01 for stability purposes.

The state given by the agent is a 7-dimensional vector given by:

- The storage at the current day, namely s_t. This quantity is normalized using $2 * \frac{s_t - 50}{500 - 50} - 1$.
- 6 basis functions $\phi_i(t)$ (with $i \in \{1, \dots 6\}$) are used to describe the time t. More specifically, $\phi_i(t) = |t - c_i|$, where $c_i = \{60, 120, 180, 240, 300, 360\}$. The agent then observes a normalized version of $\phi_i(t)$, namely $2 * \frac{\phi_i(t)}{360} - 1$.

The agent always start at day $t = 0$ with a storage s_0 of 200 and the interaction proceeds for 1080 days (i.e., 3 years). The available actions a_t are discrete and 21. Each action i represents the amount of water that the agent intends to release at

[17] As we shall show, at the beginning of the training process, with these parameters, the agent will rarely reach the goal.

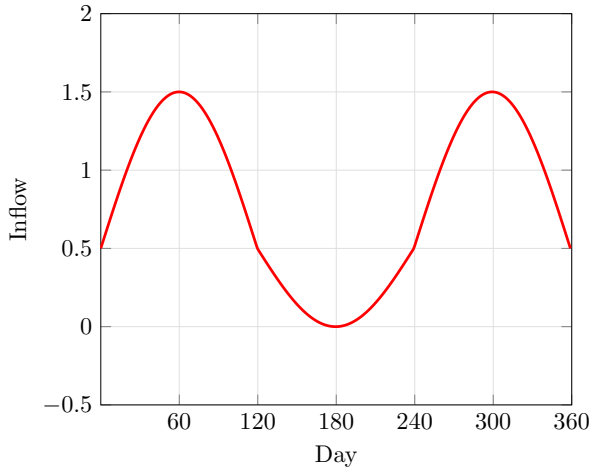

Figure 6. Mean inflow i_t per day t of the Dam control environment considered in our experiments. The inflow is considered over a period of 1 year.

day t. Differently from Tirinzoni et al. (2018), we are considering a case in which there are operational constraints on the amount of water to release (i.e., the agent cannot take all the possible values $a_t \in [0, s_t]$, but it is limited to $\{0, \dots, 20\}$). Moreover, we consider a control frequency of 3 days: once an action has been chosen at day t, it is persisted for 3 days in a row. This is mainly for performance reasons: all the policy-gradient based method that we tried were failing to learn without this additional trick.

E.1.4. REACHER

The domain is the standard one from the MuJoCo control suite (Todorov et al., 2012). We set the episode duration to $T = 200$ timesteps, with a new goal target popping up if the previous one is reached.

E.1.5. MULTI-ECHELON SUPPLY CHAIN

As originally done, we consider the optimization problem over a period of 30 days. All the details concerning this domain (e.g., demands, lead times, inventory costs, initial states, backlog costs, prices are products are sold) can be found in Hubbs et al. (2020), indeed, we rely on their publicly available repository for our experiments. We report here, however, a couple of modifications that we have taken to improve the performances. The state of the agent (i.e., a vector v with dimension 33) has been normalized according to $\frac{v}{20} - 1$. The action space, that was originally a multi-discrete space of dimension $[100, 90, 80]$, has been shrinked to $[25, 25, 25]$ to speed up the learning process. Indeed, larger action values are highly sub-optimal given the demand curve and the inventory costs.

E.2. Policy Evaluation Experiments

Figure 7 and 8 report results for our experimental evaluation setting. Figure 7 studies the on-policy evaluation problem, where we want to evaluate the random policy with data collected from the random policy itself. Figure 8, instead, focuses on the off-policy setting: we consider the problem of estimating the policy that takes a_1 with probability 0.49 and a_2 with probability 0.51 with data collected from the random policy. In both Figures, each picture compares the performance, in term of MSE, of $\tilde{\boldsymbol{m}}^*$ against the usual uniform-in-the-horizon DCS.[18] Each experiment shows the mean MSE, together with 95% confidence intervals, over 100 runs. To conduct exhaustive experimentation, we have varied both the value of Λ abd γ. The results are consistent with our theory: for small values of γ (or, equivalently, for larger values of T) the benefits of $\tilde{\boldsymbol{m}}^*$ increases.

[18]Notice that in the considered environment, the exact value of any policy can easily be computed in closed form. It follows that the MSE can be computed exactly.

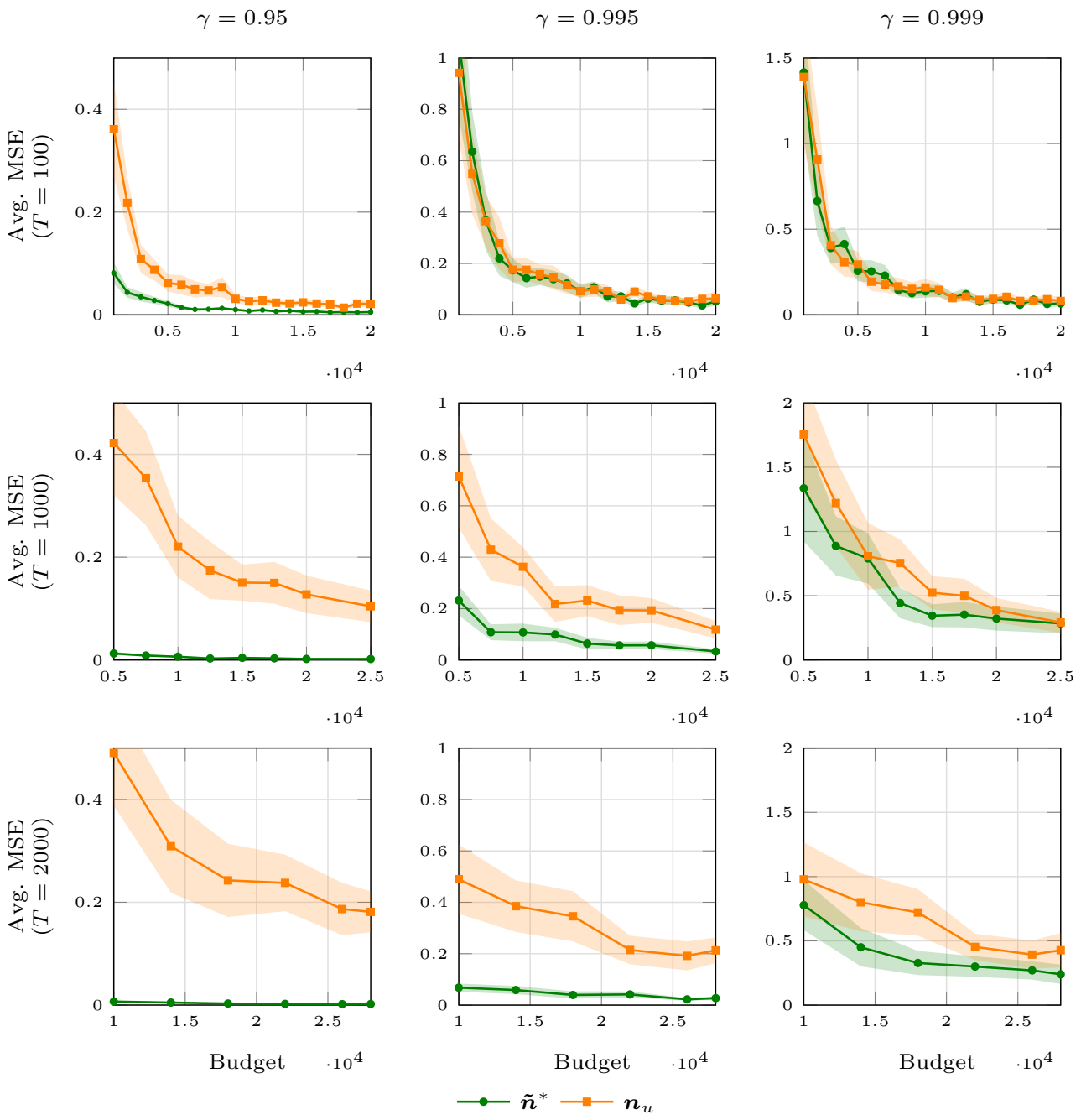

Figure 7. Experimental results (mean and 95% confidence intervals of 100 runs) on the Policy Evaluation domain described in Section E.1. Each picture reports the average MSE (i.e., y-axis) against the budget that has been spent to collect trajectories (i.e., x-axis). More specifically, we report results when using our approximately optimal DCS $\tilde{\boldsymbol{n}}^*$ and the uniform data collection strategy $\boldsymbol{n}_u$. The policy that is estimated is the uniform one, and the data have been collected on-policy. The first row of the figure is obtained with $T = 100$, the second one with $T = 1000$, and the third one with $T = 2000$. The third column with $\gamma = 0.95$, the second one with $\gamma = 0.995$, and the third one with $\gamma = 0.999$.

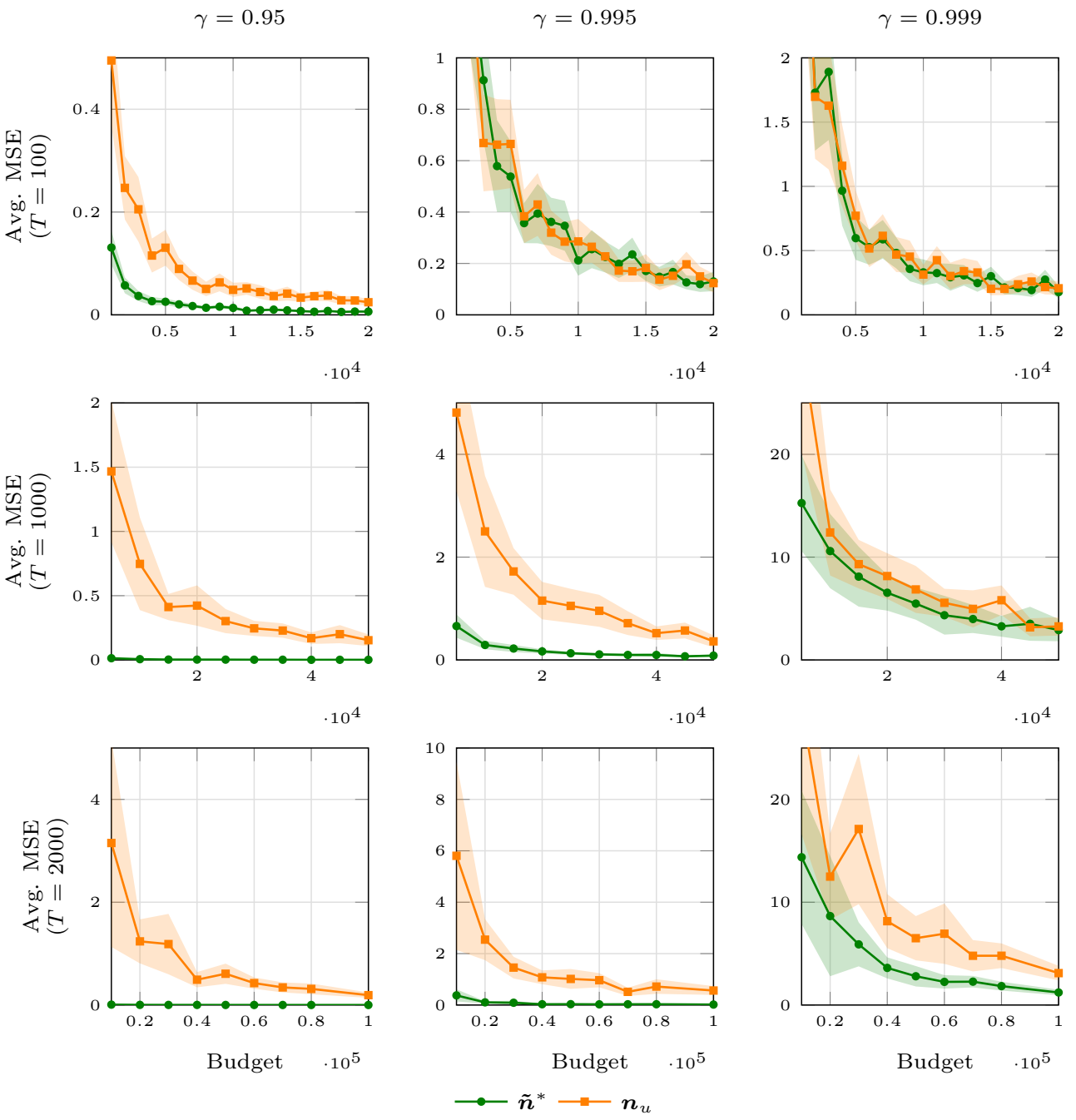

Figure 8. Experimental results (mean and 95% confidence intervals of 100 runs) on the Policy Evaluation domain described in Section E.1. Each picture reports the average MSE (i.e., y-axis) against the budget that has been spent to collect trajectories (i.e., x-axis). More specifically, we report results when using our approximately optimal DCS $\tilde{\boldsymbol{n}}^*$ and the uniform data collection strategy $\boldsymbol{n}_u$. The policy that is estimated takes a_1 with probability 0.49 and a_2 with probability 0.51; the data have been collected using the random policy (i.e., off-policy evaluation). The first row of the figure is obtained with $T = 100$, the second one with $T = 1000$, and the third one with $T = 2000$. The third column with $\gamma = 0.95$, the second one with $\gamma = 0.995$, and the third one with $\gamma = 0.999$.

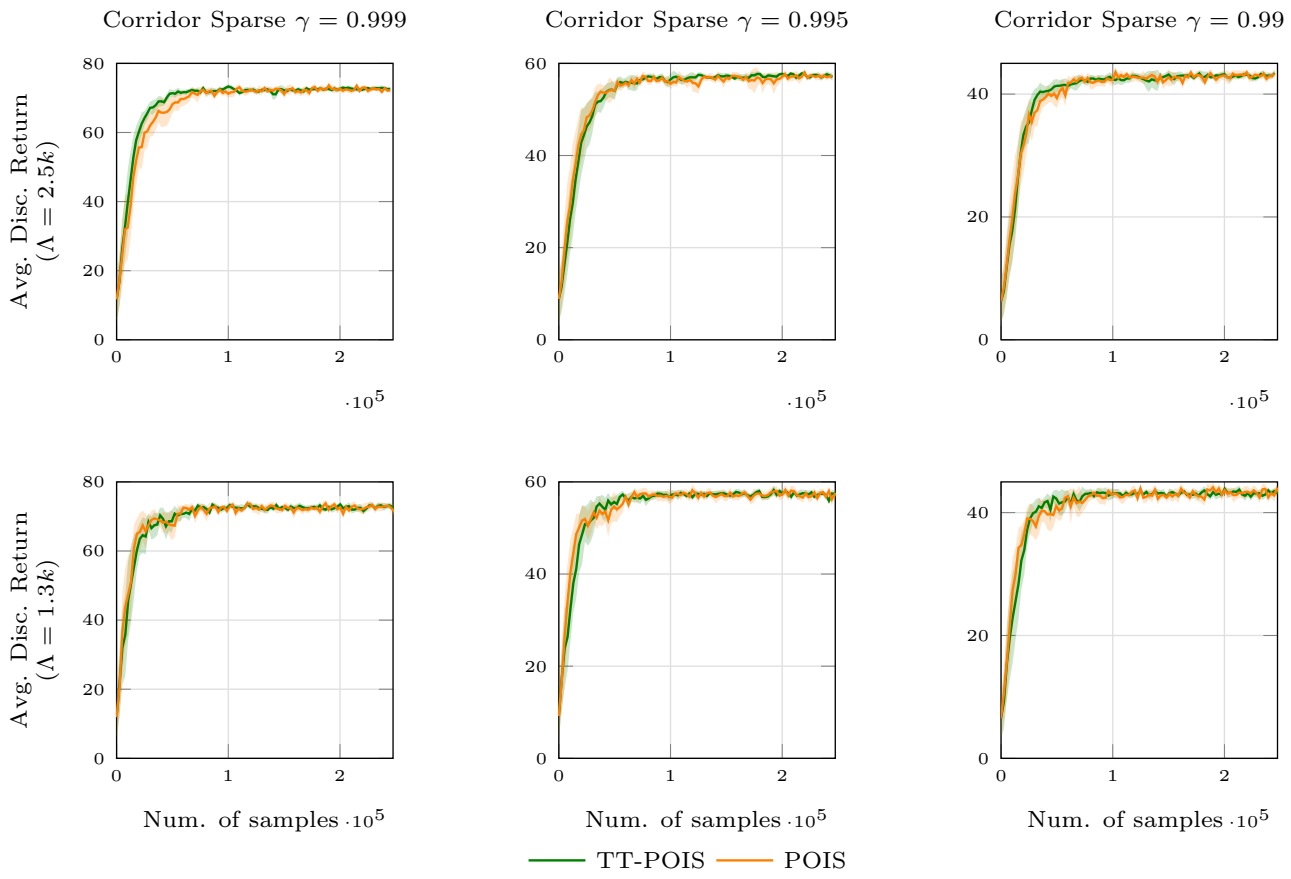

Figure 9. Experimental results (mean and 95% confidence intervals of 15 runs) on the Corridor Sparse domain with different values of γ and Λ. More specifically, the first row is obtained with $\Lambda = 2500$ and the second one with $\Lambda = 1300$. The first column is obtained training the algorithm with $\gamma = 0.999$, the second one with $\gamma = 0.995$, and the third one with $\gamma = 0.99$. The reported metric is the average of the discounted return with the corresponding value of γ.

E.3. Policy Optimization Ablations

Figure 9 and 10 report results on the Corridor domain presented in Section E.1. More specifically, Figure 9 reports the result for the sparse reward setting, while Figure 10 focuses on the dense reward one. Each plot shows the mean and the 95% confidence intervals over 15 runs of the average discounted return. To conduct exhaustive experimentation, we varied the values of γ and Λ.

Some observations are in order. First of all, from Figure 9, TT-POIS shows a robust behavior even in this sparse reward scenario. Notice that, at the beginning of the learning process, since the performance is close to 0, the agent rarely reaches the goal (i.e., it sparsely receives positive feedback from the environment). Yet, TT-POIS still obtains the same learning curves as POIS. Secondly, from Figure 10, we can appreciate a significant benefit of TT-POIS over POIS for the dense reward setting, especially with small values of γ. This is consistent with what have been highlighted in Section 5. Finally, Figure 12 and 11 report the undiscounted average return as a metric. The previous considerations extend to these results as well.

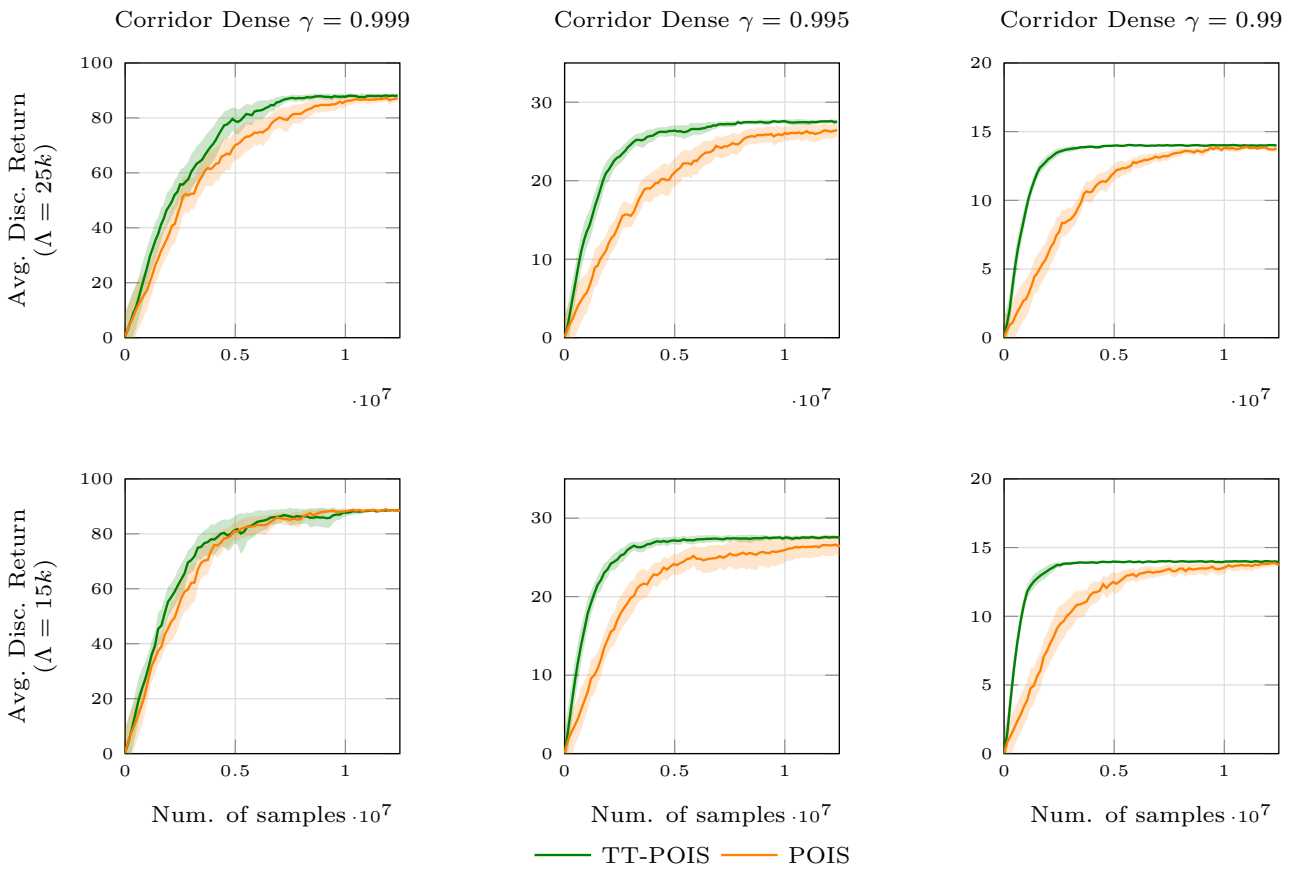

Figure 10. Experimental results (mean and 95% confidence intervals of 15 runs) on the Corridor Dense domain with different values of γ and Λ. More specifically, the first row is obtained with $\Lambda = 25000$ and the second one with $\Lambda = 15000$. The first column is obtained training the algorithm with $\gamma = 0.999$, the second one with $\gamma = 0.995$, and the third one with $\gamma = 0.99$. The reported metric is the average of the discounted return with the corresponding value of γ.

E.4. Additional Optimization Results: varying Λ and γ

In this Section, we provide additional experimental optimization results. More specifically, Figures 13, 14 and 15 reports results for the Dam, Supply Chain and Reacher environments respectively (average return over 5 run with 95% confidence intervals), while varying the value of Λ and γ. For the Dam domain we test the following combinations of Λ and γ: $\gamma \in \{0.95, 0.995, 0.999\}$ and $\Lambda \in \{4320, 8640\}$. For the Reacher domain, instead, we test $\gamma \in \{0.95, 0.995, 0.999\}$ and $\Lambda \in \{4000, 8000\}$. Finally, for the Supply Chain, we report $\gamma \in \{0.95, 0.97, 0.999\}$ and $\Lambda \in \{2400, 3900\}$.

As one can see, what has been highlighted in Section 5 replicates consistently.

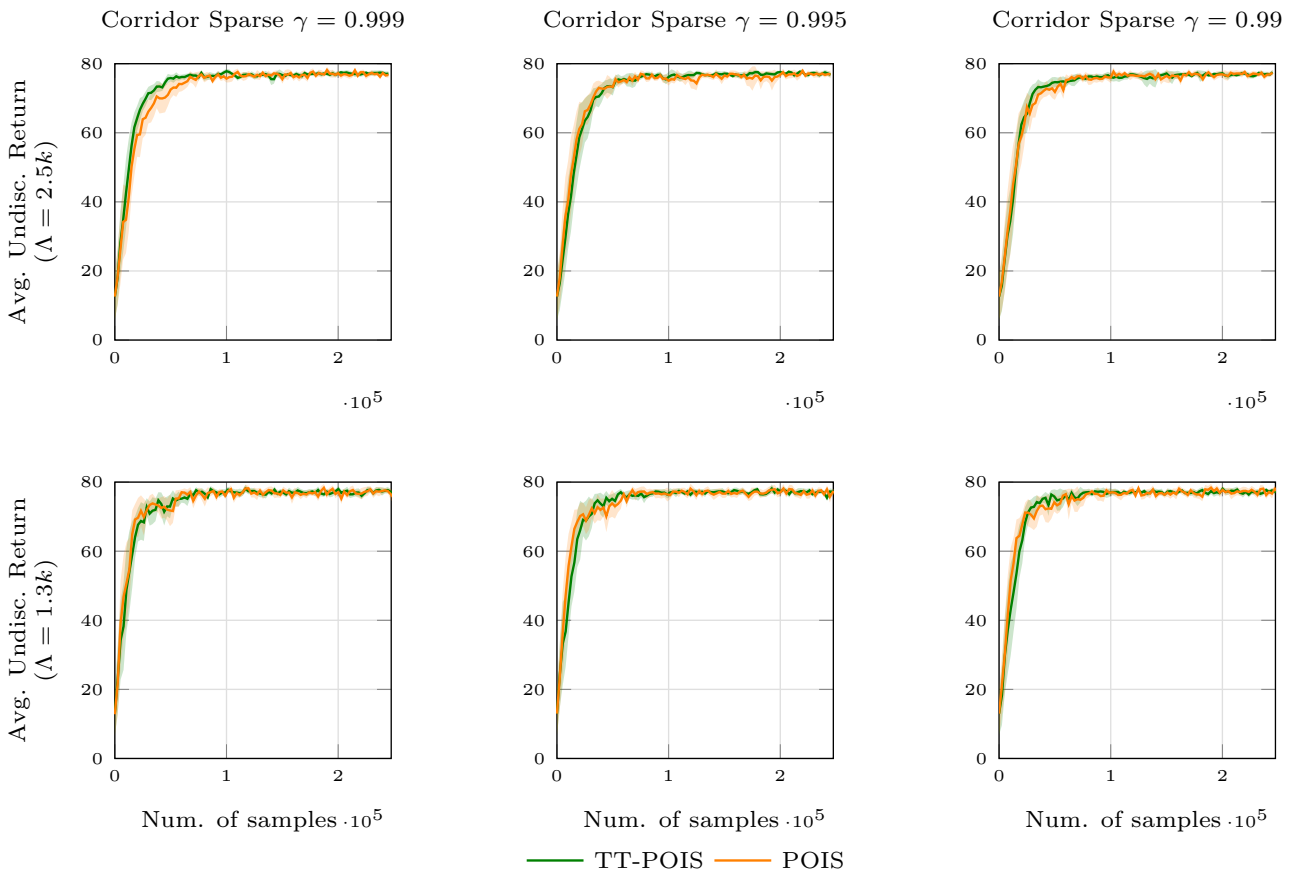

Figure 11. Experimental results (mean and 95% confidence intervals of 15 runs) on the Corridor Sparse domain with different values of γ and Λ. More specifically, the first row is obtained with $\Lambda = 2500$ and the second one with $\Lambda = 1300$. The first column is obtained with $\gamma = 0.999$, the second one with $\gamma = 0.97$, and the third one with $\gamma = 0.99$. The reported metric is the average of undiscounted return (i.e., $\gamma = 1$.)

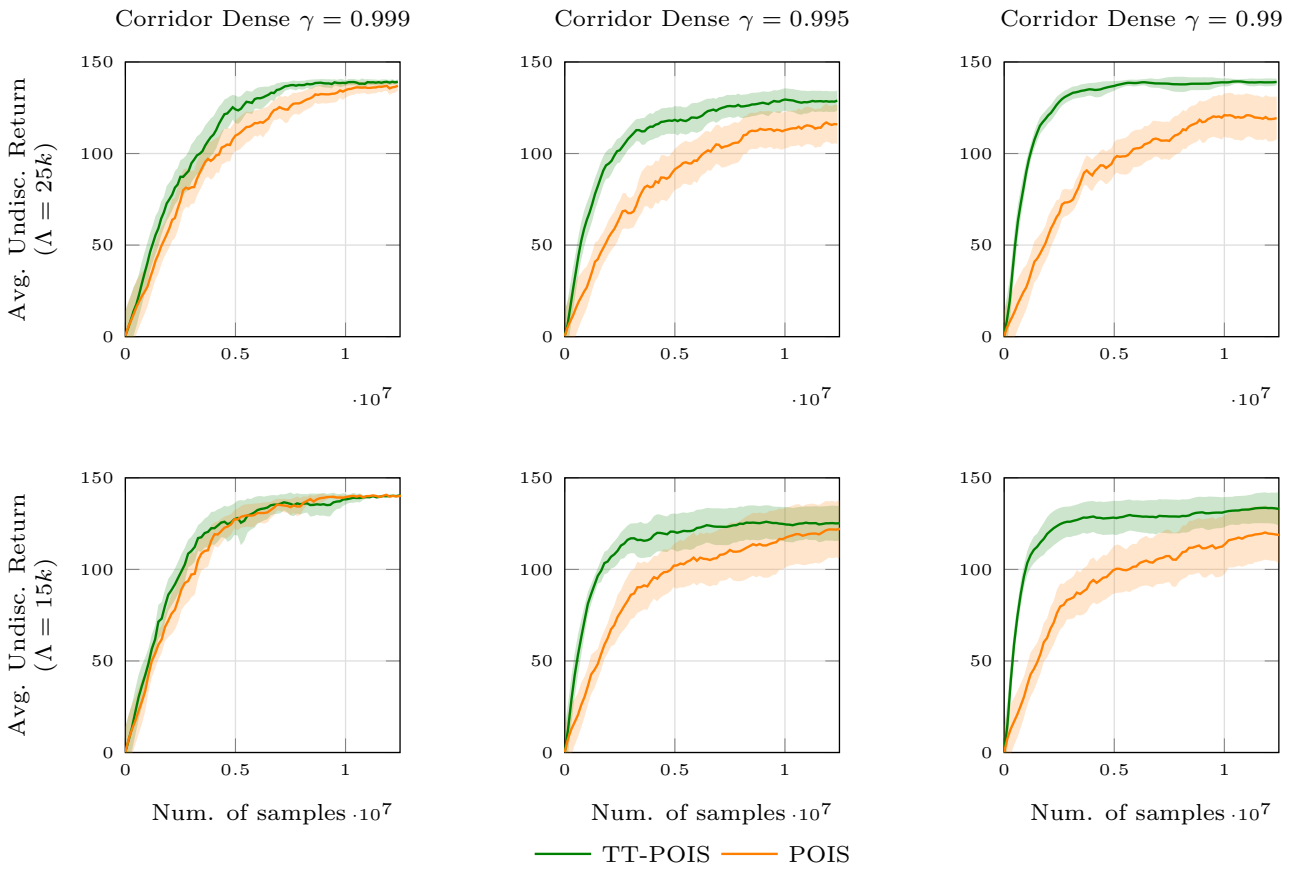

Figure 12. Experimental results (mean and 95% confidence intervals of 15 runs) on the Corridor Dense domain with different values of γ and Λ. More specifically, the first row is obtained with $\Lambda = 25000$ and the second one with $\Lambda = 15000$. The first column is obtained with $\gamma = 0.999$, the second one with $\gamma = 0.97$, and the third one with $\gamma = 0.99$. The reported metric is the average of undiscounted return (i.e., $\gamma = 1$.)

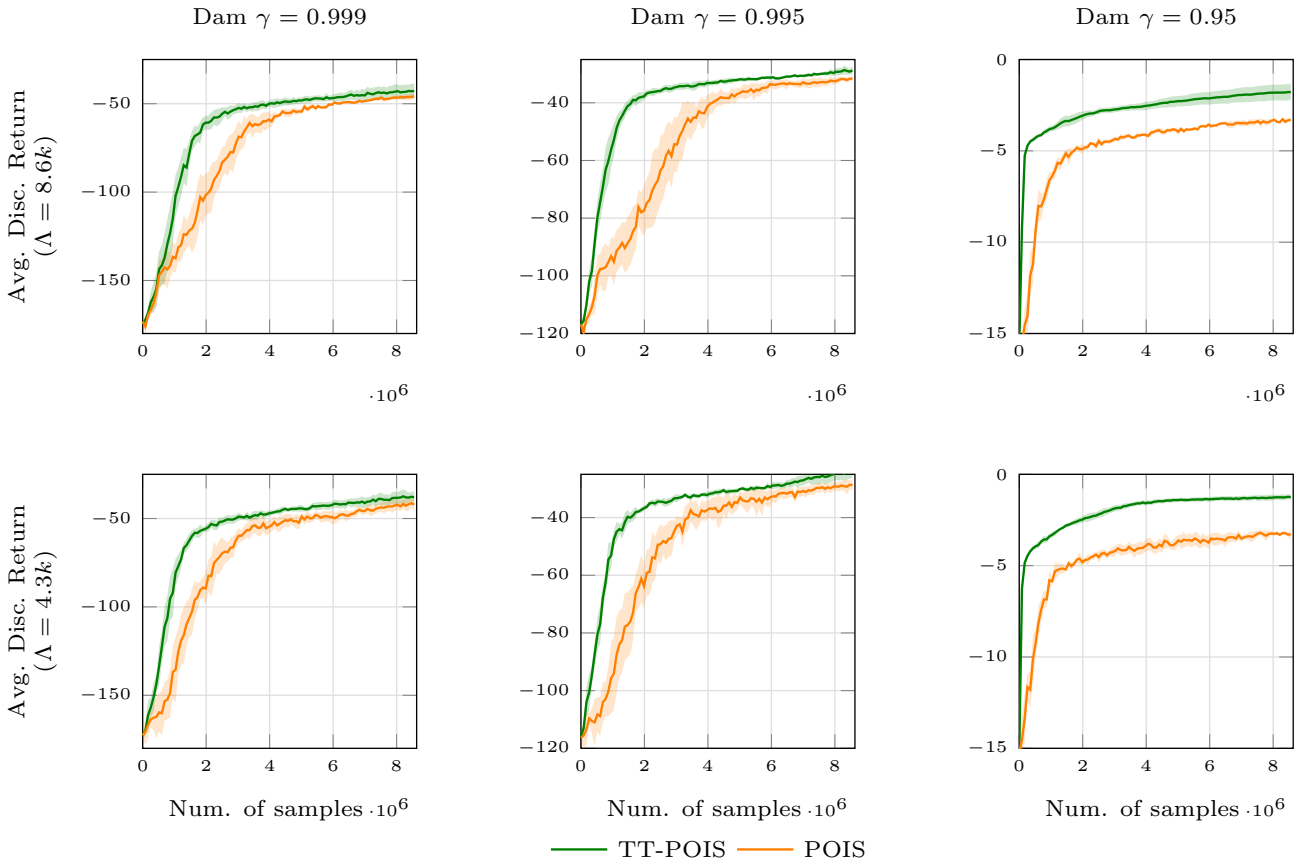

Figure 13. Experimental results (mean and 95% confidence intervals of 5 runs) on the Dam domain with different values of γ and Λ. More specifically, the first row is obtained with $\Lambda = 8640$ and the second one with $\Lambda = 4320$. The first column is obtained training the algorithm with $\gamma = 0.999$, the second one with $\gamma = 0.995$, and the third one with $\gamma = 0.95$. The reported metric is the average of the discounted return with the corresponding value of γ.

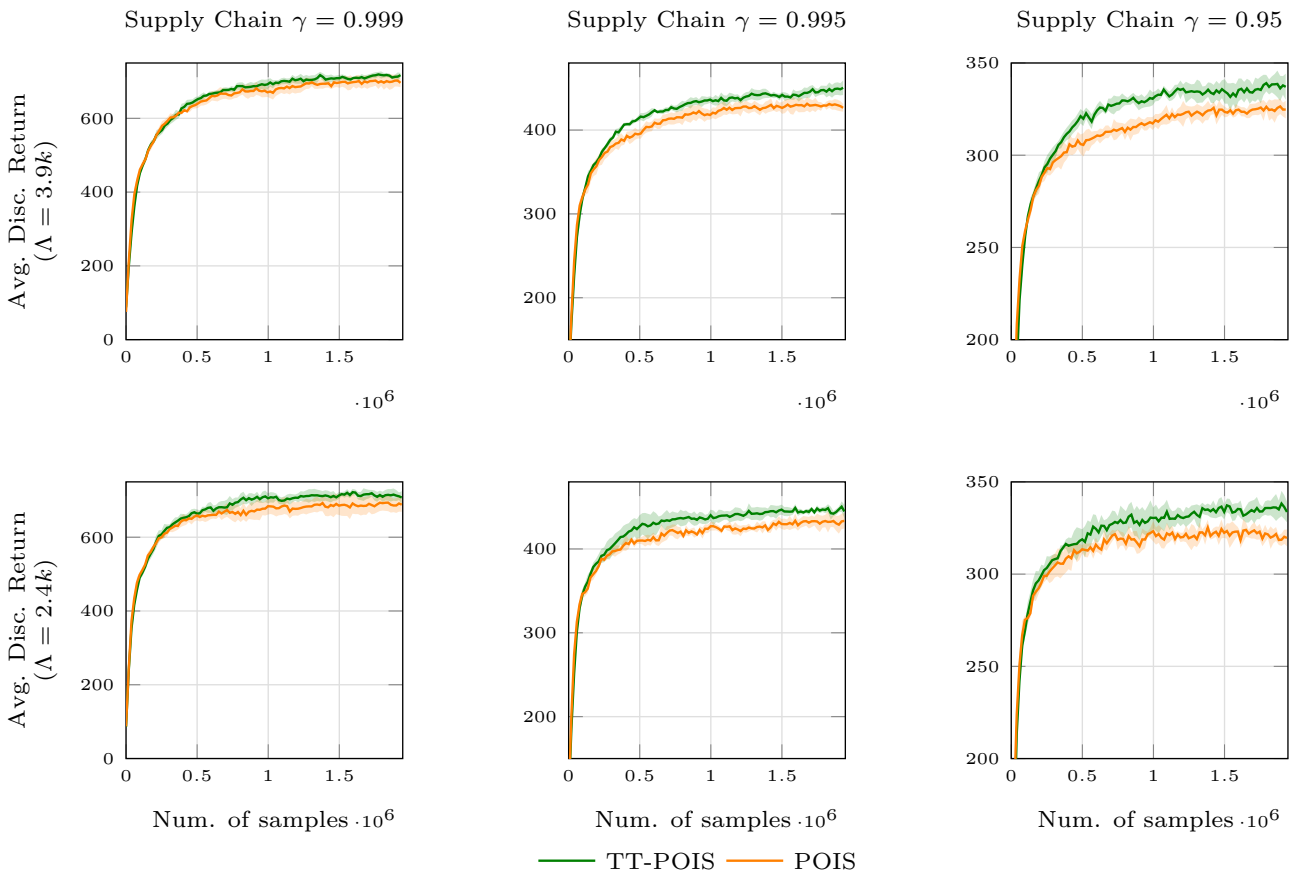

Figure 14. Experimental results (mean and 95% confidence intervals of 5 runs) on the Supply Chain domain with different values of γ and Λ. More specifically, the first row is obtained with $\Lambda = 3900$ and the second one with $\Lambda = 2400$. The first column is obtained training the algorithm with $\gamma = 0.999$, the second one with $\gamma = 0.97$, and the third one with $\gamma = 0.95$. The reported metric is the average of the discounted return with the corresponding value of γ.

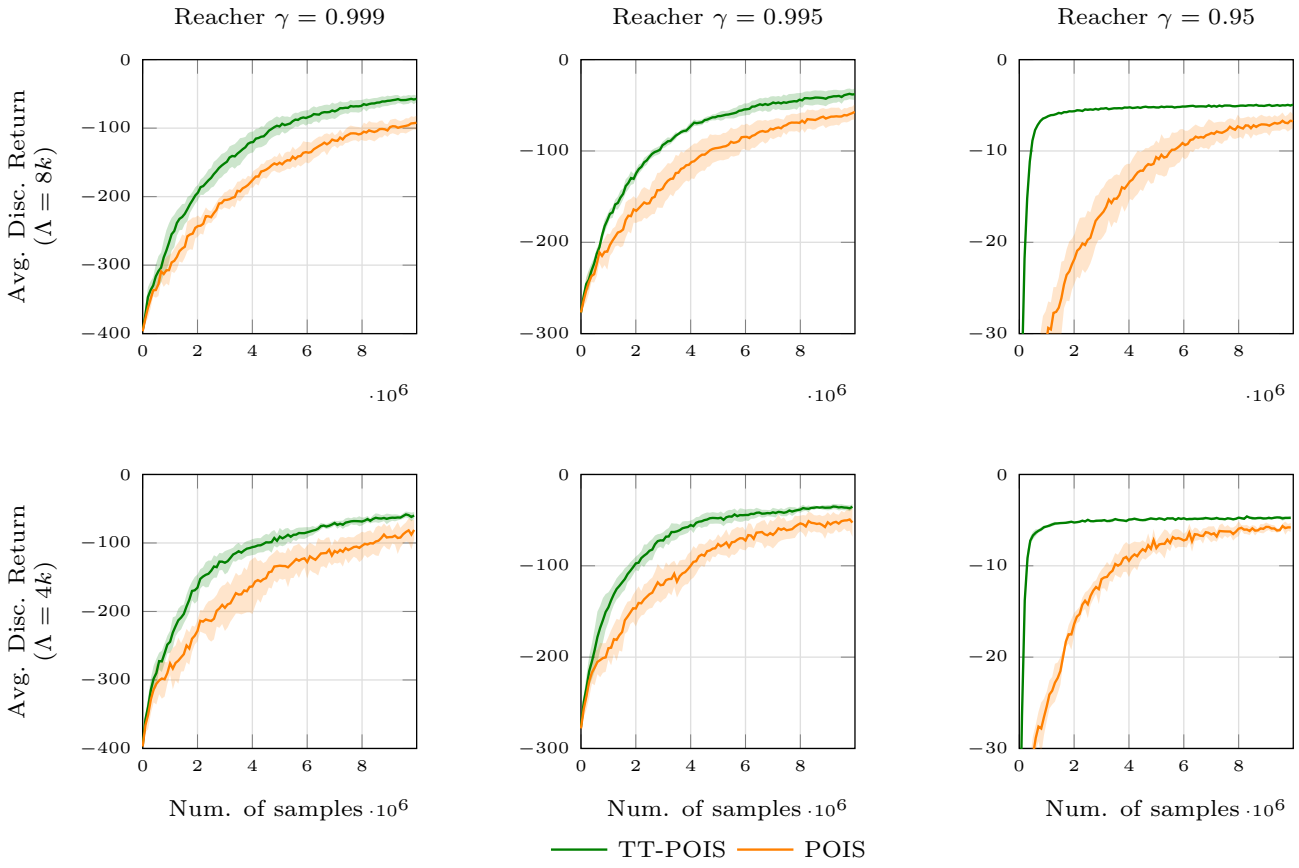

Figure 15. Experimental results (mean and 95% confidence intervals of 5 runs) on the Reacher domain with different values of γ and Λ. More specifically, the first row is obtained with $\Lambda = 8000$ and the second one with $\Lambda = 4000$. The first column is obtained training the algorithm with $\gamma = 0.999$, the second one with $\gamma = 0.995$, and the third one with $\gamma = 0.95$. The reported metric is the average of the discounted return with the corresponding value of γ.

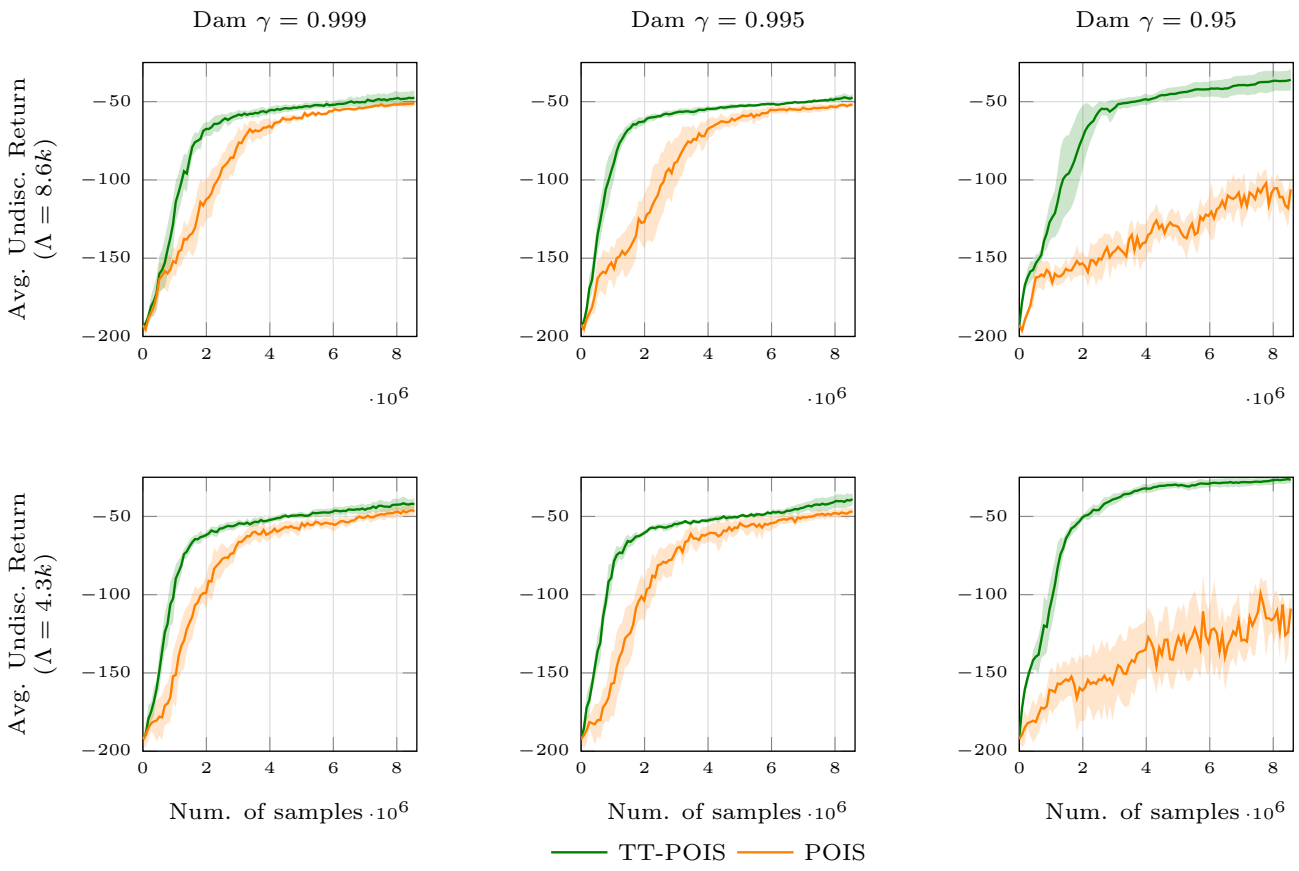

Figure 16. Experimental results (mean and 95% confidence intervals of 5 runs) on the Dam domain with different values of γ and Λ. More specifically, the first row is obtained with $\Lambda = 8640$ and the second one with $\Lambda = 4320$. The first column is obtained with $\gamma = 0.999$, the second one with $\gamma = 0.995$, and the third one with $\gamma = 0.95$. The reported metric is the average of undiscounted return (i.e., $\gamma = 1$.)

E.5. Additional Optimization Results: undiscounted performance

In this Section, we report the undiscounted average return for the experiments of Section E.4. More specifically, Figures 16, 19 and 18 reports results for the Dam, Reacher and Supply Chain environments respectively (average undiscounted return over 5 run with 95% confidence intervals).

As we can appreciate, in these scenarios, the advantages of TT-POIS over POIS replicates even for the undiscounted return metric.

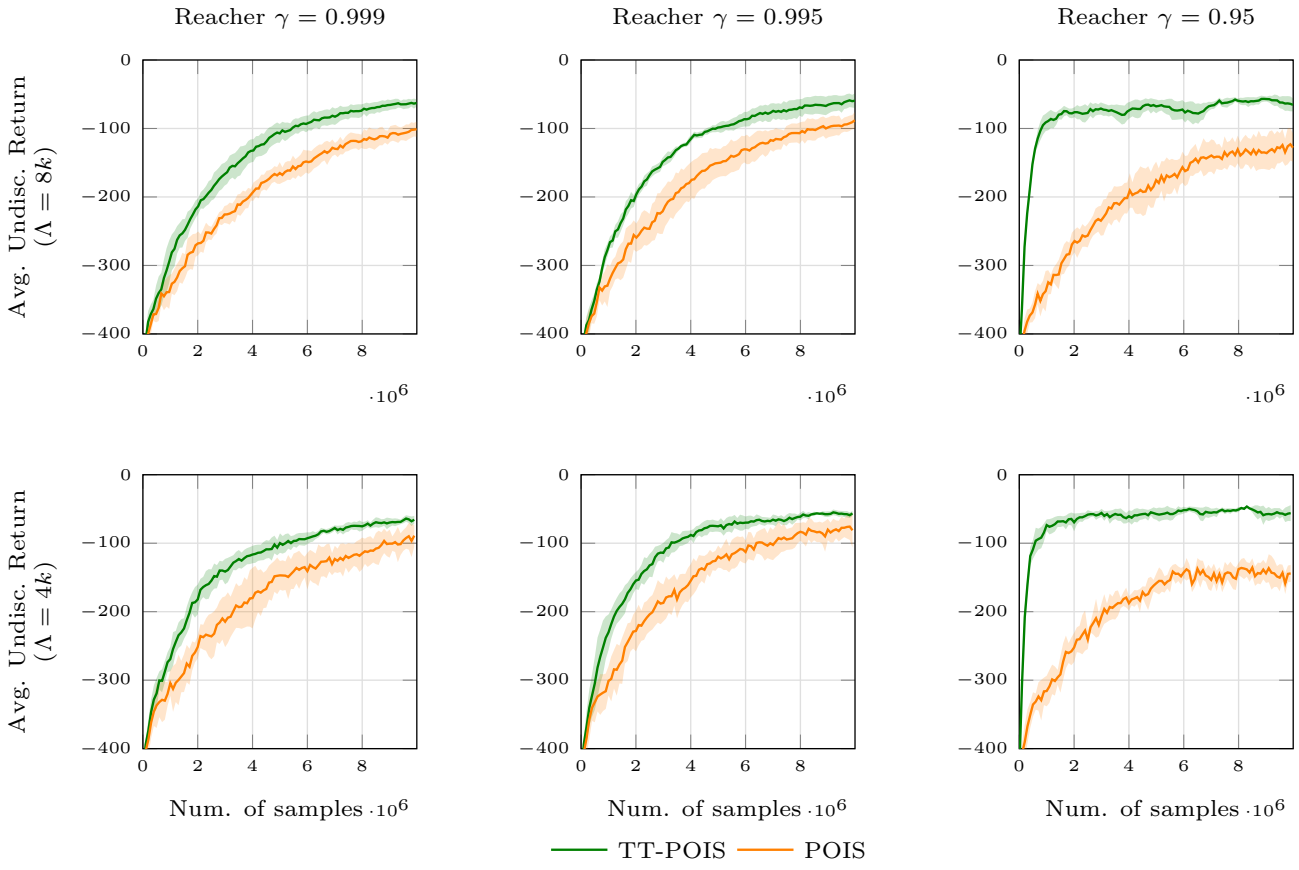

Figure 17. Experimental results (mean and 95% confidence intervals of 5 runs) on the Reacher domain with different values of γ and Λ. More specifically, the first row is obtained with $\Lambda = 8000$ and the second one with $\Lambda = 4000$. The first column is obtained with $\gamma = 0.999$, the second one with $\gamma = 0.995$, and the third one with $\gamma = 0.95$. The reported metric is the average of undiscounted return (i.e., $\gamma = 1$.)

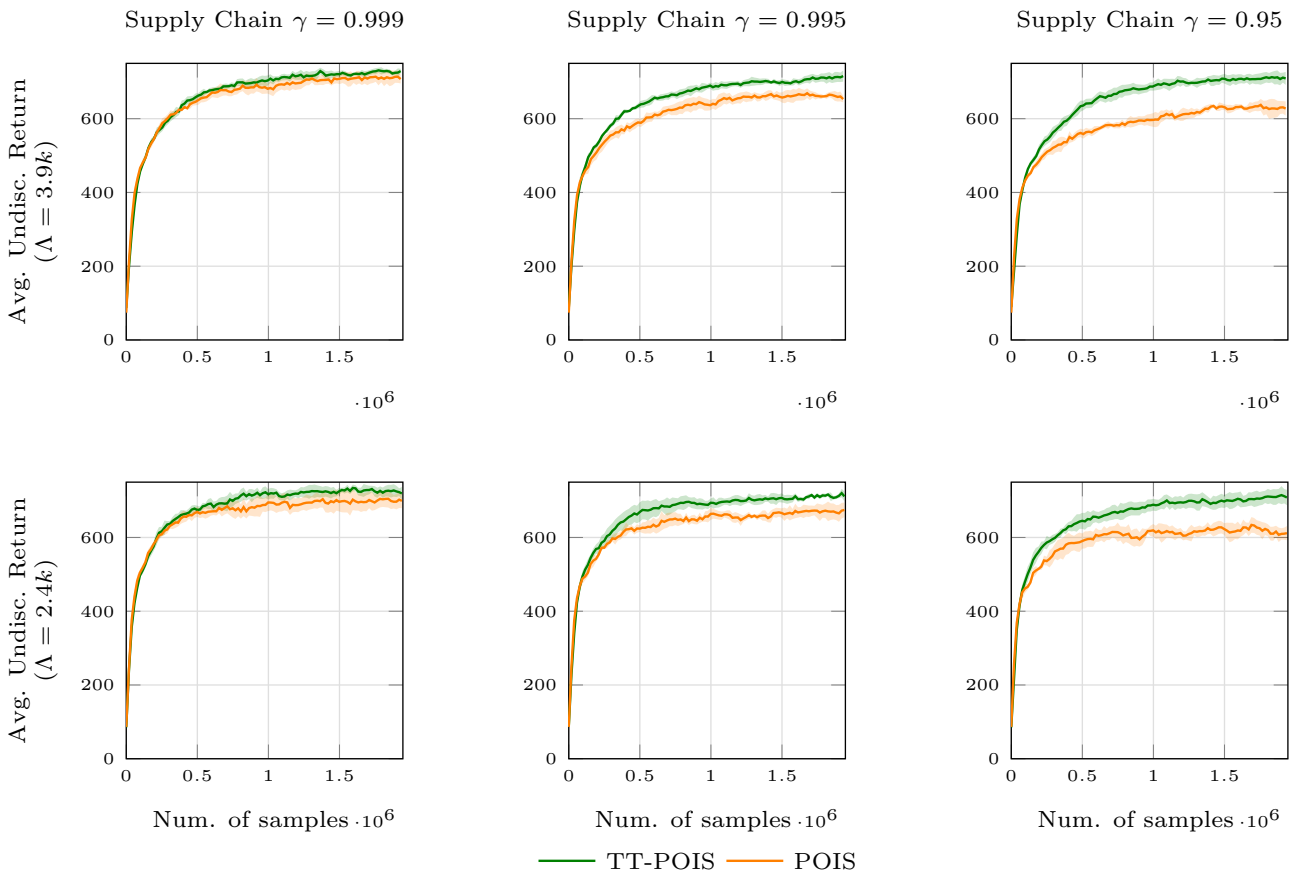

Figure 18. Experimental results (mean and 95% confidence intervals of 5 runs) on the Supply Chain domain with different values of γ and Λ. More specifically, the first row is obtained with $\Lambda = 3900$ and the second one with $\Lambda = 2400$. The first column is obtained with $\gamma = 0.999$, the second one with $\gamma = 0.97$, and the third one with $\gamma = 0.95$. The reported metric is the average of undiscounted return (i.e., $\gamma = 1$.)

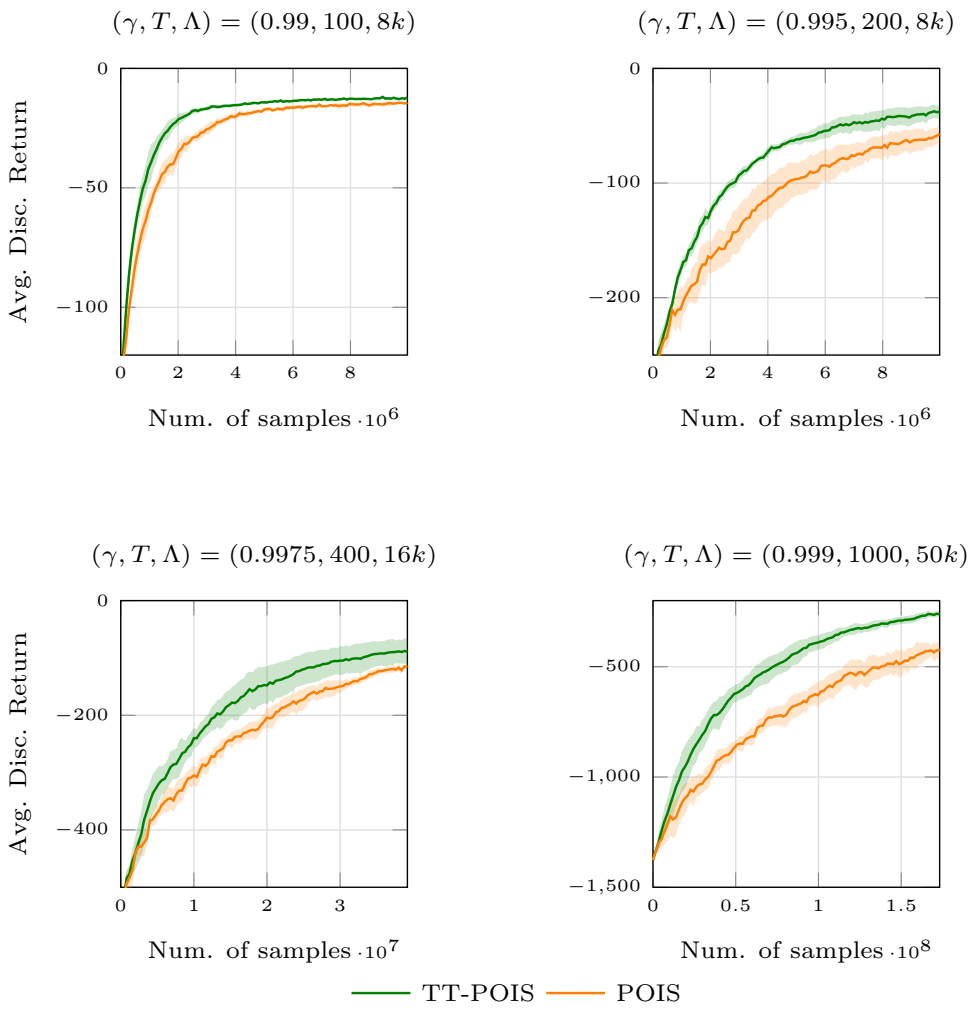

Figure 19. Experimental results (mean and 95% confidence intervals of 5 runs) on the Reacher domain varying both γ and T.

E.6. Additional Optimization Results: varying T and γ jointly

In this Section, we provide additional optimization results on the Reacher domain where we vary T and γ jointly. More specifically, we compare POIS and TT-POIS on the Reacher with $(T = 100, \gamma = 0.99)$, $(T = 200, \gamma = 0.995)$, $(T = 400, \gamma = 0.9975)$, and $(T = 1000, \gamma = 0.999)$. Notice that γ^T is roughly constant across these combinations. Furthermore, we remark for larger values of T, the optimization process is clearly significantly harder, thus requiring more training iterations and larger batch sizes to get meaningful learning results. For these reasons, for $T = 400$ we used $\Lambda = 16k$, while for $T = 1000$, we used $\Lambda = 50k$, while for $T = 100$ and $T = 200$, we used $\Lambda = 8k$ as for Figure 2.

E.7. Hyper-parameters and other details

We have run the experiments using 88 Intel(R) Xeon(R) CPU E7-8880 v4 @ 2.20GHz cpus and 94 GB of RAM.

We now provide details on the hyper-parameters that were used to generate the results. Table 1, 2, 3, 4 and 5 provides hyper-parameters for POIS and TT-POIS on the different domains. For all values of γ, we used the same hyper-parameters. The hyper-parameters on the number of offline iterations refers to Line 4 of Algorithm 1.

Table 1. Corridor Sparse Rewards Hyper-parameters for POIS and TT-POIS

Hyper-parameter	$\mathbf{\Lambda = 2500}$	$\mathbf{\Lambda = 1300}$
Neural Network Size	$[64, 32]$	$[64, 32]$
Weight Initialization	Normc	Normc
Activation Function	Xavier	Xavier
Confidence δ	0.9	0.9
Number of offline iterations	10	10
Importance Weight Clipping	100	100
$R_{\text{MIN-MAX}}$	Not applied	Not applied

Table 2. Corridor Dense Rewards Hyper-parameters for POIS and TT-POIS

Hyper-parameter	$\mathbf{\Lambda = 25000}$	$\mathbf{\Lambda = 15000}$
Neural Network Size	$[64, 32]$	$[64, 32]$
Weight Initialization	Normc	Normc
Activation Function	Xaiver	Xavier
Confidence δ	0.7	0.7
Number of offline iterations	10	10
Importance Weight Clipping	100	100
$R_{\text{MIN-MAX}}$	Not applied	Not applied

Table 3. Dam Hyper-parameters for POIS and TT-POIS

Hyper-parameter	$\mathbf{\Lambda = 8600}$	$\mathbf{\Lambda = 4320}$
Neural Network Size	$[64, 32]$	$[64, 32]$
Weight Initialization	Normc	Normc
Activation Function	Tanh	Tanh
Confidence δ	0.7	0.6
Number of offline iterations	10	10
Importance Weight Clipping	Not applied	Not applied
$R_{\text{MIN-MAX}}$	Not applied	Not applied

Table 4. Supply Chain Hyper-parameters for POIS and TT-POIS

Hyper-parameter	$\mathbf{\Lambda = 3900}$	$\mathbf{\Lambda = 2400}$
Neural Network Size	$[100, 50, 25]$	$[100, 50, 25]$
Weight Initialization	Normc	Normc
Activation Function	Tanh	Tanh
Confidence δ	0.005	0.005
Number of offline iterations	20	20
Importance Weight Clipping	100	100
$R_{\text{MIN-MAX}}$	Not applied	Not applied

Table 5. Reacher Hyper-parameters for POIS and TT-POIS

Hyper-parameter	$\mathbf{\Lambda = 8000}$	$\mathbf{\Lambda = 4000}$
Neural Network Size	$[100, 50, 25]$	$[100, 50, 25]$
Weight Initialization	Normc	Normc
Activation Function	Tanh	Tanh
Confidence δ	0.8	0.8
Number of offline iterations	20	20
Importance Weight Clipping	Not applied	Not applied
$R_{\text{MIN-MAX}}$	5	5