Truncating Trajectories in Monte Carlo Policy Evaluation: an Adaptive Approach

Riccardo Poiani
DEIB, Politecnico di Milano
riccardo.poiani@polimi.it

Nicole Nobili
DEIB, Politecnico di Milano
nicole.nobili@mail.polimi.it

Alberto Maria Metelli
DEIB, Politecnico di Milano
albertomaria.metelli@polimi.it

Marcello Restelli
DEIB, Politecnico di Milano
marcello.restelli@polimi.it

Abstract

Policy evaluation via Monte Carlo (MC) simulation is at the core of many MC Reinforcement Learning (RL) algorithms (e.g., policy gradient methods). In this context, the designer of the learning system specifies an interaction budget that the agent usually spends by collecting trajectories of *fixed length* within a simulator. However, is this data collection strategy the best option? To answer this question, in this paper, we consider as quality index the variance of an unbiased policy return estimator that uses trajectories of different lengths, i.e., *truncated*. We first derive a closed-form expression of this variance that clearly shows the sub-optimality of the fixed-length trajectory schedule. Furthermore, it suggests that adaptive data collection strategies that spend the available budget sequentially might be able to allocate a larger portion of transitions in timesteps in which more accurate sampling is required to reduce the variance of the final estimate. Building on these findings, we present an *adaptive* algorithm called **R**obust and **I**terative **D**ata collection strategy **O**ptimization (RIDO). The main intuition behind RIDO is to split the available interaction budget into mini-batches. At each round, the agent determines the most convenient schedule of trajectories that minimizes an empirical and robust estimate of the estimator's variance. After discussing the theoretical properties of our method, we conclude by assessing its performance across multiple domains. Our results show that RIDO can adapt its trajectory schedule toward timesteps where more sampling is required to increase the quality of the final estimation.

1 Introduction

In Reinforcement Learning [RL, Sutton and Barto, 2018], an agent acts in an unknown, or partially known, environment to maximize/estimate the infinite expected discounted sum of an external reward signal, i.e., the expected return. Monte Carlo evaluation [MC, Owen, 2013] is at the core of many successful RL algorithms. Whenever a simulator with reset possibility is available to the learning systems designer, a large family of approaches [Williams, 1992, Baxter and Bartlett, 2001, Schulman et al., 2015, 2017, Cobbe et al., 2021] that can be used to solve the RL problem relies on MC simulations for estimating performance or gradient estimates on the task to be solved. In this scenario, since the goal is to estimate the expected infinite sum of rewards, the designer usually specifies a sufficiently large estimation horizon T, along with a transition budget $\Lambda = QT$, so that the agent

16th European Workshop on Reinforcement Learning (EWRL 2023).

interacts with the simulator, via MC simulation, collecting a batch of Q episodes of length T.[1] In this sense, the agent spends its available budget Λ *uniformly along the estimation horizon*.

In the context of MC policy evaluation, where the goal lies in estimating the performance of a given policy via MC simulations, Poiani et al. [2023] have recently shown that, given the discounted nature of the RL objective, this uniform-in-the-horizon budget allocation strategy may not be the best option. The core intuition behind their work is that, since rewards are exponentially discounted through time, early interactions weigh exponentially more than late ones, and, consequently, a larger portion of the available budget Λ should be dedicated to estimating the initial rewards. To theoretically validate this point, the authors designed a *non-adaptive* budget allocation strategy which, by exploiting the reset possibility of the simulator, leads to the collection of trajectories of different lengths, i.e., *truncated*. They show that this approach provably minimizes Höeffding-like confidence intervals [Boucheron et al., 2003] around the empirical estimates of the expected return. Remarkably, this implies a robustness w.r.t. the uniform strategy that holds for any pair of environment and policy to be evaluated, thus, clearly establishing the theoretical benefits of the proposed method.

Nevertheless, it has to be noticed that although minimizing confidence intervals around the expected return estimator comes with desirable theoretical guarantees (e.g., PAC-bound improvements [Even-Dar et al., 2002]), the resulting schedule of trajectories is computed *before* the interaction with the environment (being determined by the discount factor). Consequently, as the usual uniform-in-the-horizon scheme, it fails to adapt to the peculiarities of the problem at hand, and, ultimately, might not produce a low error estimate. For the sake of clarity, we illustrate this sub-optimality of pre-determined schedule of trajectories with the following extreme examples.

Example 1. *Consider an environment where the reward is gathered only at the end of the horizon T (e.g., a goal-based). In this scenario, any strategy that truncates trajectories is intuitively sub-optimal, and we expect that an intelligent agent will spend all its budget according to the uniform schedule.*

Example 2. *Conversely, consider a problem where the reward is different from 0 in the first interaction step only (e.g., in the case of a highly sub-optimal policy that immediately reaches the "zero reward region" of an environment); the uniform schedule wastes a significant portion of its budget collecting samples without variability, and, to reduce the estimation error, we would like the agent to spend all of its interaction budget estimating the reward of the first action.*

Abstracting away from the previous examples, we realize that the main issue of existing approaches arises from the fact that determining a schedule of trajectories *before* interacting with the environment does not allow the agent to adapt it to the environment peculiarities, allocating more samples where this is required to obtain a high-quality estimate. For this reason, in this work, we focus on designing *adaptive* data collection strategies that aim directly at minimizing the *error* of the final estimate. Our main intuition lies in splitting the available budget Λ into mini-batches and adapting *online* the data collection strategy of the agent based on the previously collected information.

Original Contributions and Outline After introducing the necessary notation and backgrounds (Section 2), we consider the problem of maximizing the estimation quality of a policy expected return estimator using trajectories of different lengths collected via MC simulation with a finite budget Λ of transitions (Section 3). More specifically, since we use an *unbiased* return estimator, we consider its variance as a quality index, of which we derive a closed-form expression and analyze it for every possible schedule of trajectories. Then, we define the optimal trajectories schedule as the one that attains the minimum variance subject to the available budget constraint. As expected, computing this optimal data collection strategy requires knowledge of the underlying environment (e.g., the variance of the rewards at each timestep), which is not available to the agent prior to the interaction. Nevertheless, as we shall see, all the quantities that define the optimal strategy can be estimated from the data. These facts confirm our intuition about the weakness of non-adaptive schedules of trajectories and suggest that algorithms that spend the available budget Λ iteratively might be able to dynamically allocate their budget to minimize the variance of the final estimate. Building on these findings, in Section 4, we present a novel algorithm, **R**obust and **I**terative **D**ata collection strategy **O**ptimization (RIDO), which splits its available budget Λ into mini-batches of interactions that are allocated sequentially to minimize an empirical and robust estimate of the

[1] While another large class of RL algorithms is based on Temporal Difference [TD, Sutton and Barto, 2018] learning, which do not require the finite horizon nor the reset possibility, Monte Carlo simulation approaches continue to be extensively adopted. Indeed, unlike TD methods, they can be applied effortlessly to non-Markovian environments, which is a common occurrence in real-world problems.

objective function of interest, i.e., the variance of the estimator. Furthermore, we perform a statistical analysis on the behavior of RIDO, and we derive theoretical guarantees expressed as upper bounds on the variance of the policy return estimator. Our result shows that, under favorable conditions, the variance of the return estimator computed by RIDO is of the same order as the one of the oracle's baseline. To conclude, in Section 5, we conduct an experimental comparison between RIDO and non-adaptive schedules. As we verify, our method achieves the most competitive performance across different domains, discount factor values, and budget, thus clearly highlighting the benefits of adaptive strategies over pre-determined ones.

2 Backgrounds and Notation

This section provides the notation and necessary backgrounds used in the rest of this document.

Markov Decision Processes A discrete-time Markov Decision Process [MDP, Puterman, 2014] is defined as a tuple $\mathcal{M} \coloneqq (\mathcal{S}, \mathcal{A}, R, P, \gamma, \nu)$, 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 the specifies the reward $R(s, a)$ received by the agent upon taking action a in state s, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$[2] is the transition kernel that specifies the probability distribution over the next states $P(\cdot|s, a)$, when taking action a in state s, $\gamma \in (0, 1)$ is the discount factor, and $\nu \in \Delta(\mathcal{S})$ is the distribution over initial states. The agent's behavior is modeled by a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which for each state s, prescribes a distribution over actions $\pi(\cdot|s)$. A trajectory $\boldsymbol{\tau}_h$ of length h is a sequence of states and actions $(s_0, a_0, s_1, \dots s_{h-1}, a_{h-1}, s_h)$ observed by following π for h steps, where $s_0 \sim \nu$, $a_t \sim \pi(\cdot|s_t)$, and $s_{t+1} \sim P(\cdot|s_t, a_t)$ for $t < h$. The return of a trajectory is defined as $G(\boldsymbol{\tau}_h) = \sum_{t=0}^{h-1} \gamma^t R_t$, where R_t is shortcut for $R(s_t, a_t)$. The agent that is following policy π is evaluated according to expected cumulative discounted sum of rewards over an estimation horizon T,[3], namely $J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t R_t\right]$, where the expectation is taken w.r.t. the stochasticity of the policy, the transition kernel, and the initial state distribution.

Data Collection Strategy Poiani et al. [2023] formalized the concept of Data Collection Strategy (DCS) to model how the agent collects data within an environment. More specifically, given an interaction budget $\Lambda \in \mathbb{N}$ such that $\Lambda \mod T = 0$, a DCS is defined as a T-dimensional vector $\boldsymbol{m} \coloneqq (m_1, \dots, m_T)$ where $m_h \in \mathbb{N}$ and $\sum_{h=1}^{T} m_h h = \Lambda$. Each element m_h specifies the number of trajectories of length h that the agent collects in the environment while following a policy π. Given a DCS $\boldsymbol{m}$, it is possible to compute the total number of steps $\boldsymbol{n} \coloneqq (n_0, \dots, n_{T-1})$ that will be gathered by the agent at any step t; more specifically, the following relationship holds: $n_{T-1} = m_T$, and $n_t = n_{t+1} + m_{t+1}$ for $t < T - 1$. For this reason, in the rest of the paper we will adopt the most convenient symbol depending on the context. For any DCS $\boldsymbol{m}$ such that $m_T \geq 1$ holds, it is possible to build the following unbiased estimator of $J(\pi)$:

$$\hat{J}_{\boldsymbol{m}}(\pi) = \sum_{h=1}^{T} \sum_{i=1}^{m_h} \sum_{t=0}^{T-1} \frac{\gamma^t}{n_t} R_t^{(i)}. \tag{1}$$

The two external summations in Equation (1) sum over the collected trajectories of different lengths a rescaled empirical trajectory return, where the reward at step t is divided by the number of samples collected at step t.[4] In the case the budget Λ is spent uniformly, i.e., $\boldsymbol{m} = \left(0, \dots, 0, \frac{\Lambda}{T}\right)$, Equation (1) reduces to the usual Monte Carlo estimator of $J(\pi)$, namely $\frac{T}{\Lambda} \sum_{i=1}^{\Lambda/T} \sum_{t=0}^{T-1} \gamma^t R_t^{(i)}$.

Robust Data Collection Strategy Optimization Leveraging the estimator of Equation (2), Poiani et al. [2023] investigated alternatives to the usual uniform-in-the-horizon DCS from the worst-case perspective of confidence intervals [Boucheron et al., 2003]. More specifically, given $\boldsymbol{m}$ such that $m_T \geq 1$, the estimator of Equation (1) enjoys the following generalization of the Höeffding

[2] Given a set $\mathcal{X}$, we denote with $\Delta(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$.

[3] As common in Monte-Carlo simulation [see e.g., Papini et al., 2022] we approximate the infinite horizon MDP model with a finite estimation horizon T. Indeed, if T is sufficiently large, i.e., $T = \mathcal{O}\left(\frac{1}{1-\gamma} \log \frac{1}{\epsilon}\right)$, the expected return computed with horizon T is ϵ close to the infinite-horizon one [Kakade, 2003].

[4] Rescaling by n_t prevents the estimate from being biased towards time steps for which more samples are available.

confidence intervals holding with probability at least $1-\delta$:

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

where $d_t = \frac{\gamma^t(\gamma^t+\gamma^{t+1}-2\gamma^T)}{1-\gamma}$ controls the relative importance of samples gathered at step t. Poiani et al. [2023] designed a closed-form DCS that provably minimizes the bound of Equation (2). Since d_t is a decreasing function of time whose decay speed is governed by the discount factor γ, the aforementioned DCS gives priority to the collection of experience at earlier time steps, i.e., it truncates the trajectories. Note that the smaller γ, the higher the number of samples reserved for earlier time steps. We refer the reader for Theorem 3.3 and Theorem B.10 of their work for the exact expressions of the resulting robust DCS. However, we remark that the resulting schedule is non-adaptive (i.e., it is computed before the interaction with the environment takes place) and its shape depends exclusively on Λ, γ, and T.

3 Toward Adaptive Data Collection Strategies

In this section, we lay down the theoretical groundings behind optimizing data collection strategies that directly aim at minimizing the final estimation error. We stick to methods that adopt the estimator of Equation (1), which has a simple interpretation and desirable theoretical properties. More specifically, since the estimator is unbiased, the Mean Squared Error (MSE) simply reduces to the variance. For this reason, to set a proper baseline for DCS optimization (i.e., an optimal strategy according to the MSE), we first analyze the variance of the estimator of Equation (1) for an arbitrary DCS $\boldsymbol{m}$. The following Theorem (proof in Appendix B) summarizes our result.

Theorem 3.1. *Consider a generic DCS $\boldsymbol{m}$ such that $m_T \geq 1$, then:*

$$\mathbb{V}\mathrm{ar}_{\boldsymbol{m}}\left[\hat{J}_{\boldsymbol{m}}(\pi)\right] = \sum_{t=0}^{T-1}\frac{1}{n_t}\left(\gamma^{2t}\mathbb{V}\mathrm{ar}(R_t) + 2\sum_{t'=t+1}^{T-1}\gamma^{t+t'}\mathbb{C}\mathrm{ov}(R_t, R_{t'})\right) =: \sum_{t=0}^{T-1}\frac{f_t}{n_t}. \tag{3}$$

Theorem 3.1 expresses, in closed form, the variance of the estimator of Equation (1) when adopted with an arbitrary DCS that guarantees the estimation to be unbiased (i.e., $m_T \geq 1$). From Equation (3), we can see that this variance results in a summation, over the different time steps, of $\frac{1}{n_t}$, i.e., the reciprocal of the number of samples collected under $\boldsymbol{m}$ at step t, multiplied by the variance of the reward at step t plus the covariances between R_t and the rewards gathered at future steps. For brevity, we shortcut this term with f_t. Furthermore, Theorem 3.1 leads to a direct formulation of an optimal DCS baseline for our setting. More specifically, given a budget Λ, we define the optimal DCS $\boldsymbol{n}^*$ for the estimator in Equation (1) as the solution of the following optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{n}} \quad & \sum_{t=0}^{T-1}\frac{1}{n_t}\left(\gamma^{2t}\mathbb{V}\mathrm{ar}(R_t) + 2\sum_{t'=t+1}^{T-1}\gamma^{t+t'}\mathbb{C}\mathrm{ov}(R_t, R_{t'})\right) \\
\text{s.t.} \quad & \sum_{t=0}^{T-1} n_t \leq \Lambda \\
& n_t \geq n_{t+1}, \quad \forall t \in \{0, \dots, T-2\} \\
& n_t \in \mathbb{N}_+, \quad \forall t \in \{0, \dots, T-1\},
\end{aligned} \tag{4}$$

where the constraints $n_t \geq n_{t+1}$ directly encode the sequential nature of the interaction with the environment. At this point, some comments are in order. First of all, we notice that the above optimization problem nicely captures the intuitive examples of Section 1.

Example 1 (*cont.*)**.** *When the reward is different from* 0 *in the last interaction step only, the objective function reduces to* $\frac{\gamma^{2(T-1)}}{n_{T-1}}\mathbb{V}\mathrm{ar}\left[R_{T-1}\right]$*, which is clearly minimized for the uniform strategy.*

Example 2 (*cont.*)**.** *Conversely, when the reward is different from* 0 *in the first step only, we obtain* $\frac{\mathbb{V}\mathrm{ar}[R_0]}{n_0}$*, meaning that the entire interaction budget should be dedicated to estimate* R_0*.*

Algorithm 1 Robust and Iterative DCS Optimization (RIDO).

Require: Interaction budget Λ, batch size b, robustness level β, policy π

1: Collect $\mathcal{D}$ using policy π and $\hat{\boldsymbol{n}}_0 = \left(\frac{b}{T}, \dots, \frac{b}{T}\right)$

2: Set $K = \frac{\Lambda}{b}$ and initialize empirical estimates $\sqrt{\widehat{\mathbb{V}}\mathrm{ar}_1\left[R_t\right]}$ and $\widehat{\mathbb{C}}\mathrm{ov}_1\left[R_t, R_t'\right]$

3: **for** $i = 1, \dots, K-1$ **do**

4: Collect $\mathcal{D}_i$ using policy π and $\hat{\boldsymbol{n}}_i$, where $\hat{\boldsymbol{n}}_i$ is computed solving problem (5)

5: Update empirical estimates $\sqrt{\widehat{\mathbb{V}}\mathrm{ar}_i\left[R_t\right]}$ and $\widehat{\mathbb{C}}\mathrm{ov}_i\left[R_t, R_t'\right]$ using $\mathcal{D}_i$ and set $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_i$

6: **end for**

Furthermore, the formulation of problem (4) highlights the pitfalls of a non-adaptive data collection strategies. Indeed, consider, for the sake of clarity, the two examples mentioned above. Before executing policy π, the agent has no way of distinguishing between the two different objective functions, i.e., $\frac{\gamma^{2(T-1)}}{n_{T-1}}\mathbb{V}\mathrm{ar}\left[R_{T-1}\right]$ and $\frac{\mathbb{V}\mathrm{ar}[R_0]}{n_0}$, and, consequently, any pre-determined schedule fails to adapt to the actual objective function. More generally, this is because the optimal strategy resulting from the optimization problem (4) can be computed prior to the interaction with the environment only by an oracle that knows in advance the underlying reward process induced by the agent's policy π in the MDP. Nevertheless, we note that all the terms appearing in the objective function, i.e., the only unknowns in optimization problem (4), can be estimated if some interactions with the environment are available to the agent. This suggests that strategies that sequentially allocate the available budget Λ might successfully adapt their DCS to minimize Equation (3), i.e., the variance of the return estimator.

4 Robust and Iterative DCS Optimization

Given the findings of Section 3, we now present our algorithmic solution that aims at avoiding the highlighted pitfalls of pre-determined DCSs. Our approach is called **R**obust and **I**terative **D**ata collection strategy **O**ptimization (RIDO), and its pseudocode is available in Algorithm 1. The central intuition behind RIDO lies in splitting the available budget Λ into mini-batches of interactions that the agent will allocate *sequentially*. At each iteration, the agent will compute the most convenient schedule of trajectories that optimizes an *empirical and robust* version of the objective function presented in (4), whose quality improves as the agent gathers more data.

We now describe in-depth the behavior of the algorithm. For simplicity of exposition and analysis, we suppose that the size of the mini-batch b is such that $b \mod T = 0$ and $b \geq 2T$. At the beginning (Lines 1-2 in Algorithm 1), the agent spends the first mini-batch $\hat{\boldsymbol{n}}_{\mathbf{0}}$ at collecting $\frac{b}{T}$ trajectories of length T (i.e., the uniform approach). This preliminary collection phase is a starting round in which some initial experience is gathered to properly initialize estimates of relevant quantities used throughout the algorithm. More specifically, at each iteration i, the agent maintains empirical estimates of the unknown quantities that define the variance of the estimate, i.e., the standard deviation of the reward at step t, namely $\sqrt{\widehat{\mathbb{V}}\mathrm{ar}_i\left[R_t\right]}$, and the covariances between rewards at different steps, namely $\widehat{\mathbb{C}}\mathrm{ov}_i\left(R_t, R_{t'}\right)$. Then, at each round (Lines 4-5 in Algorithm 1), the DCS of the current mini-batch $\hat{\boldsymbol{n}}_{\boldsymbol{i}}$ is computed solving the optimization problem (5) whose objective function is a robust estimate of the objective function of the original optimization problem (4). More specifically, at each round i, the agent aims at solving the following problem:

$$
\begin{aligned}
\min_{\boldsymbol{n}} \quad & \sum_{t=0}^{T-1} \frac{1}{n_t}\left(\gamma^{2t}\left(\sqrt{\widehat{\mathbb{V}}\mathrm{ar}_i(R_t)} + \mathbf{C}^{\sigma}_{i,t}\right)^2 + 2\sum_{t'=t+1}^{T-1}\gamma^{t+t'}\left(\widehat{\mathbb{C}}\mathrm{ov}_i(R_t, R_{t'}) + \mathbf{C}^{c}_{i,t,t'}\right)\right) \\
\text{s.t.} \quad & \sum_{t=0}^{T-1} n_t \leq b \\
& n_t \geq n_{t+1}, \quad \forall t \in \{0, \dots, T-2\} \\
& n_t \in \mathbb{N}_+, \quad \forall t \in \{0, \dots, T-1\},
\end{aligned}
\tag{5}
$$

where $\mathrm{C}^{\sigma}_{i,t}$ and $\mathrm{C}^{c}_{i,t,t'}$ are exploration bonuses for variances and covariances respectively, defined as:

$$\mathrm{C}^{\sigma}_{i,t} := \sqrt{\frac{2\log(\beta)}{\sum_{j=1}^{i-1}\hat{n}_{j,t}}}, \qquad \mathrm{C}^{c}_{i,t,t'} := 3\sqrt{\frac{2\log(\beta)}{\sum_{j=1}^{i-1}\hat{n}_{j,t'}}}, \tag{6}$$

where $\beta \geq 1$ is a hyper-parameter that specifies the amount of exploration used to solve the optimization problem, and $\hat{n}_{j,t}$ is the number of samples collected by RIDO during phase j at time step t. We now provide further explanations on the optimization problem (5) and Equation (6). First of all, we notice how each term in the original objective function, namely f_t , is replaced with its relative empirical estimation plus exploration bonuses, each of which is directly related to components within f_t, e.g., $\mathbb{V}\mathrm{ar}\left(R_t\right)$ is replaced with $\left(\sqrt{\widehat{\mathbb{V}}\mathrm{ar}_i(R_t)} + \mathrm{C}^{\sigma}_{i,t}\right)^2$ and $\mathbb{C}\mathrm{ov}\left(R_t, R_{t'}\right)$ is replaced with $\widehat{\mathbb{C}}\mathrm{ov}_i\left(R_t, R_{t'}\right) + \mathrm{C}^{c}_{i,t,t'}$. Intuitively, the purpose of the exploration bonus is to consider the uncertainty that arises from replacing exact quantities with their empirical estimation. This introduces in RIDO a source of robustness w.r.t. the noise that is intrinsically present in the underlying estimation process. At this point, concerning the shape of Equation (6), focus for the sake of exposition on $\mathrm{C}^{\sigma}_{i,t}$. First of all, we notice that the hyper-parameter β governs the robustness which is taken into account while replacing $\mathbb{V}\mathrm{ar}\left(R_t\right)$ with its empirical estimate. Larger values of β, correspond, indeed, to larger $\mathrm{C}^{\sigma}_{i,t}$, and, consequently, a higher level of robustness w.r.t. the uncertainty. Furthermore, as we can notice, $\mathrm{C}^{\sigma}_{i,t}$ decreases with the number of samples collected in the previous iterations at step t, i.e., $\sum_{j=1}^{i-1}\hat{n}_{j,t}$. This quantity coincides with the number of samples that are used to estimate $\sqrt{\widehat{\mathbb{V}}\mathrm{ar}_i\left[R_t\right]}$.[5] This formulation captures the following aspect: more data is available to the agent to estimate $\mathbb{V}\mathrm{ar}\left(R_t\right)$, the more accurate its estimate will be, and, consequently, its exploration bonus will shrink to 0. As one can expect, with this approach, the quality of the objective function used in RIDO increases with the number of iterations. Consequently, the agent will progressively adapt the mini-batch DCS toward time steps where more data is required to minimize Equation (3), i.e., the variance of the return estimator. We conclude with two remarks. First, we notice that RIDO can be applied with $\gamma = 1$, as it does not deeply rely on the property of discounted sums. Secondly, the optimization problem (5) is a complex integer and non-linear optimization problem. Before diving into the statistical analysis of RIDO, we discuss how to solve (5) in the next section.

4.1 Solving the Empirical Optimization Problem

As noticed above, directly solving problem (5) requires significant effort since it is an integer, non-linear optimization problem. In this section, we discuss how to overcome these challenges.

We first perform a *continuous relaxation*, replacing the integer constraint $n_t \in \mathbb{N}_+$ with $n_t \geq 1$. Once a solution $\bar{\boldsymbol{n}}^*$ to the relaxed optimization problem is found, it is possible to obtain a proper (i.e., integer) DCS by flooring each $\bar{n}^*_t$ and allocating the remaining budget uniformly. As we shall see, this approximation introduces constant terms in the theoretical guarantees of RIDO only. At this point, the resulting optimization problem is a non-linear problem that, unfortunately, is generally non-convex. This issue occurs when the following condition is verified for some time step t:

$$f_t = \gamma^{2t}\left(\sqrt{\widehat{\mathbb{V}}\mathrm{ar}_i(R_t)} + \mathrm{C}^{\sigma}_{i,t}\right)^2 + 2\sum_{t'=t+1}^{T-1}\gamma^{t+t'}\left(\widehat{\mathbb{C}}\mathrm{ov}_i(R_t, R_{t'}) + \mathrm{C}^{c}_{i,t,t'}\right) < 0. \tag{7}$$

To solve this challenge and make RIDO computationally efficient, we develop an approach based on a hidden property of the original optimization problem (4). More specifically, we start by noticing that even the continuous relaxation of (4) is non-convex since $f_{\bar{t}} < 0$ might occur, for some $\bar{t} \in \{0, \dots, T-2\}$, in the presence of negative covariances with future steps. In this case, however, since $\sum_{t=\bar{t}}^{T-1} f_t$ represents a proper variance, which is always non-negative, there always exists $t' > \bar{t}$ such that $\sum_{t=\bar{t}}^{t'} f_t \geq 0$. Furthermore, it is possible to show that the optimal solution of the relaxed optimization problem is uniform in the interval $\{\bar{t}, \dots, t'\}$, namely $n^*_{\bar{t}} = n^*_{\bar{t}+1} = \dots = n^*_{t'}$ (proof in

[5] Similar comments apply to $\mathrm{C}^{c}_{i,t,t'}$ as well. The only difference stands in the fact that to estimate the empirical covariance between two subsequence steps t and t', samples up to time t' are required. For this reason, the denominator implies the summation of the number of samples gathered at t' over the previous iterations.

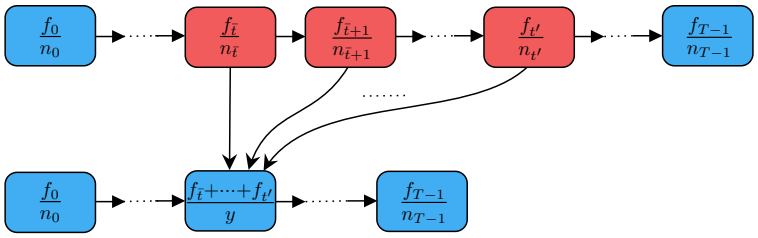

Figure 1: Visualization of the transformation between the optimization problems. The first row shows the objective function of the original optimization problem, while the second one its transformation.

Appendix B). For this reason, it is possible to define a *transformation* of the optimization problem that preserves the optimal solution, in which the variables $n_{\bar{t}}, \dots, n_{t'}$ are replaced with a single variable y. The objective function is modified accordingly, namely $\frac{f_{\bar{t}}}{n_{\bar{t}}} + \cdots + \frac{f_{t'}}{n_{t'}}$ is replaced with $\frac{f_{\bar{t}}+\cdots+f_{t'}}{y}$ in the objective function. A visualization of the transformation is proposed in Figure 1. By repeating the procedure for all the negative f_t, we obtain a transformation of the original problem which is now convex. Once the solution to this convex transformed optimization problem is found, one can quickly recover the relaxed DCS in its T-dimensional form.

Building on these results, we apply in RIDO a similar procedure that transforms the relaxed version of (5) into a new problem where the negative time steps (i.e., steps in which Equation (7) holds) are "grouped" with future time steps as long as the total summation is positive. In this way, (i) the resulting optimization problem is convex and (ii) as our analysis will reveal, this procedure has no impact on the theoretical properties of RIDO (i.e., the result is the same as assuming access to an oracle that can solve non-linear and non-convex problems). As a concluding remark, we refer the reader to Appendix B for a formal description of the above-mentioned procedure.

4.2 Theoretical Analysis

We now present theoretical guarantees on the performance of RIDO. More specifically, we derive high-probability guarantees on the variance of the estimator of Equation (1) when used with the data collected by RIDO. Before diving into the presentation of our results, we highlight some critical challenges behind the result. First, in our analysis, we do not assume access to an oracle that solves (5), but we consider the modifications discussed in Section 4.1 that make the computation tractable. This introduces a first level of challenges in the analysis (e.g., dealing with the roundings that arise from the relaxation and the peculiar strategy that overcomes the non-convexity of the optimization problem). Secondly, we notice that none of the optimization problems (4) and (5), and the ones obtained by relaxing the integer constraints, admit a closed-form solution (further details are provided in Appendix B). This clearly results in an additional challenge in our analysis. At this point, we are ready to state our main theoretical result (proof in Appendix B).

Theorem 4.1. *Let $\boldsymbol{n}^*$ be the optimal solution of problem* (4)*, f_t as in Equation* (3)*, $b \geq 2T$ and $\beta = \frac{6(T+T^2)\Lambda K}{\delta}$. Consider the DCS $\hat{\boldsymbol{n}}$ computed by Algorithm 1. Then, with probability at least $1-\delta$ it holds that:*

$$\mathbb{V}\mathrm{ar}_{\hat{\boldsymbol{n}}}\left[\hat{J}_{\hat{\boldsymbol{n}}}(\pi)\right] \leq 192\left(\frac{b}{\Lambda}\right)^{\frac{3}{2}}\log(\beta)\left(\sum_{t=0}^{T-1}\gamma^t\right)^2 + 4\mathbb{V}\mathrm{ar}_{\boldsymbol{n}^*}\left[\hat{J}_{\boldsymbol{n}^*}(\pi)\right] + \frac{2b}{\Lambda}\sum_{t:f_t<0}|f_t|. \quad (8)$$

Equation (8) comprises three terms which we now discuss in detail. The former is directly responsible for taking into account the cumulative error computed during each phase i. This term shrinks to zero with rate $\widetilde{\mathcal{O}}\left(\left(\frac{b}{\Lambda}\right)^{\frac{3}{2}}\log\Lambda\right)$. We notice that this gets smaller as we decrease b, thus suggesting to use small batch sizes. This should come as no surprise; indeed, using smaller batch sizes intuitively improves the adaptiveness of the algorithm, since a larger portion of the budget Λ will be allocated following more precise estimates of the quantities of interest. In this sense, there exists a trade-off between theoretical guarantees and computational requirements, since the number of iterations (and, thus, the number of optimization problems to be solved) grows linearly as the batch size decreases. The second term, instead, is the variance of the optimal DCS computed as in (4), and shrinks with a

rate that is at most $\frac{4}{\Lambda}\sum_{t=0}^{T-1} f_t$. This term, as we shall show below, represents a particularly desirable property. Finally, the last component of Equation (8) is related to the negative terms possibly present in the objective function, and, among the three terms, it is the one with the worst dependence on Λ. Currently, we are unsure whether this term is an artifact of the analysis, a sub-optimality of the algorithm, or a key challenge of the setting. We leave closing this gap to future work. At this point, we highlight a particular relevant property of Equation (8). Suppose that $f_t \geq 0$ holds for all time steps t (so that the last component is not present). In this case, under the mild assumption that $\sum_{t=0}^{T-1} f_t > 0$ (i.e., the variance is different from 0), for sufficiently large budget of Λ, we have that (formal statement and proof in Appendix B):

$$\mathbb{V}\mathrm{ar}_{\hat{\boldsymbol{n}}}\left[\hat{J}_{\hat{\boldsymbol{n}}}(\pi)\right] \leq 5\mathbb{V}\mathrm{ar}_{\boldsymbol{n}^*}\left[\hat{J}_{\boldsymbol{n}^*}(\pi)\right]. \tag{9}$$

Thus, in this scenario, the variance of the returned DCS computed by RIDO is proportional to the optimal one. Note that this sort of result is not possible for the uniform strategy, nor for the robust one of Poiani et al. [2023]. Further details on this point are provided in Appendix B.

5 Numerical Validation

In this section, we propose numerical validations that aim at assessing the empirical performance of RIDO. More specifically, we focus on the comparison between our approach, the classical uniform-in-the-horizon strategy, and the robust DCS by Poiani et al. [2023]. We report the results across multiple domains, values of budget Λ, and discount factor γ. As a performance index, all experiments measure the empirical variance of the estimator in Equation (1) at the end of the data collection process. Before discussing our results in detail, we describe our experimental settings in depth.

Experimental Setting In our experiments, we consider the following four domains. We start with the Inverted Pendulum [Brockman et al., 2016], a classic continuous control benchmark, where the agents' goal is to swing up a suspended body and keep it in the vertical direction. We, then, continue with the Linear Quadratic Gaussian Regulator [LQG, Curtain, 1997], where the agent controls a linear dynamical system with the objective of reducing a total cost that is expressed as a quadratic function. Then, we consider a 2D continuous navigation problem, where an agent starts at the bottom left corner of a room and needs to reach a goal region in the upper right corner. The agent receives reward 0 everywhere except inside the goal area, where the reward is positive and sampled from a Gaussian distribution. Finally, we consider the Ant environment from the MuJoCo [Todorov et al., 2012] suite, where the agent controls a four-legged 3D robot with the goal of moving it forward. Further domain details are provided in Appendix C. Concerning the policy that we evaluate for the Inverted Pendulum and the Ant, we rely on pre-trained deep RL agents made publicly available by Raffin [2020]. For the LQG, instead, we evaluate the optimal policy that is available in closed form by solving the Riccati equations, and, finally, for the 2D navigation task, we roll out a hand-designed policy that minimizes the distance of the agent's position w.r.t. to the center of the goal region. Regarding the performance index, as already anticipated, we report the variance of the empirical policy return at the end of the data collection process. Given a budget and a DCS, for a single run, we estimate this empirical variance using 100 simulations. We then average the results over 100 runs and report the empirical mean together with 95% confidence intervals. We notice that for each considered value of Λ, the experiment is repeated (i.e., we do not use data collected with smaller Λ's). To conclude, we refer the reader to Appendix C for further details on the experiments (e.g., ablations, additional results, experiments with $\gamma = 1$, hyper-parameters, visualizations of the resulting DCSs).

Results Figure 2 reports the results varying the discount factor and the available budget. The second row is obtained under the same experimental setting as the first one, but with lower values of γ. Let us first focus on the sub-optimality of the non-adaptive DCSs (i.e., the uniform strategy and the robust one of Poiani et al. [2023]). Indeed, as suggested by Theorem 3.1, being computed prior to the interaction with the environment, these algorithms cannot adapt the collection of samples to minimize the variance of the return estimator. This is clear by looking, for instance, at the results of Continuous Navigation and the LQG. Indeed, in the Continuous Navigation domain, the reward is sparse and received close to the end of the estimation horizon T. In this scenario, the robust DCS blindly truncates trajectories, thus, avoiding the collection of experience in the most relevant timesteps. Conversely, in the LQG experiments, the optimal policy that arises from the Riccati equation pays a

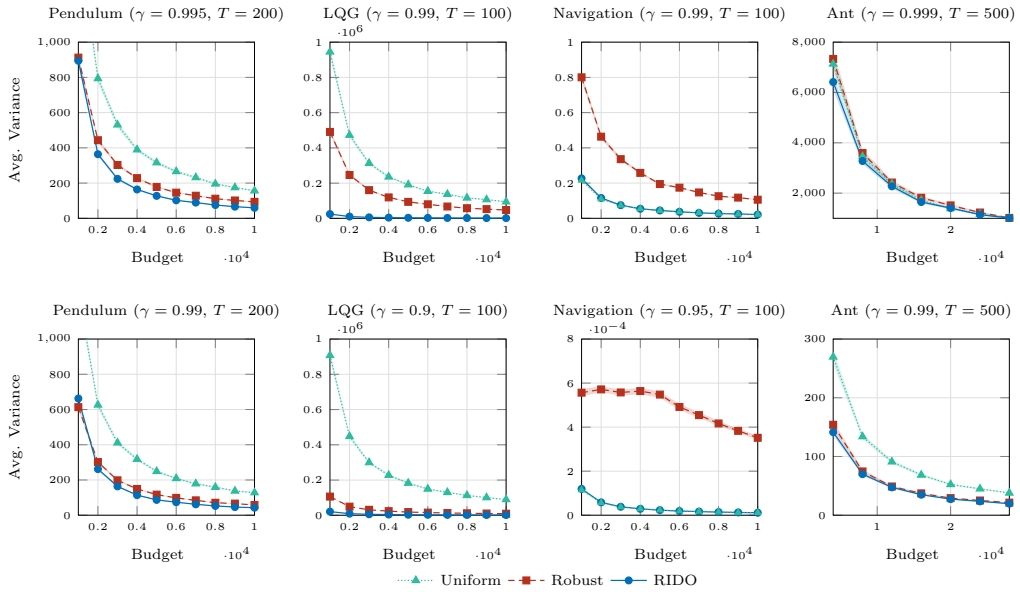

Figure 2: Empirical variance (mean and 95% confidence intervals over 100 runs) on the considered domains and baselines. The first row considers higher values of γ w.r.t. the second one.

stochastic control cost[6] at the beginning of the estimation horizon to bring the state of the system close to stability, after which the reward will remain almost constant. In this case, the uniform DCS results in a highly sub-optimal behavior as most of the estimation uncertainty is related to the initial interaction steps. RIDO, on the other hand, thanks to its adaptivity, is able to obtain the best results in both domains. Indeed, in the Continuous Navigation problem, it achieves the same performance level as the uniform strategy, while in the LQG it even outperforms the robust DCS of Poiani et al. [2023]. The reason is that Poiani et al. [2023] truncates trajectories solely depending on the value of γ, and, therefore, it might waste a portion of its budget in trajectories of sub-optimal length, while RIDO, since it aims at minimizing the variance of the final estimation, is able to focus the collection of data in the most convenient way. Similar comments to those made for the LQG hold for the Pendulum domain as well. Concerning the Ant environment, instead, we notice that for $\gamma = 0.999$ there is no significant difference between any of the presented schedules. Interestingly, however, as soon as we decrease γ to 0.99, we can appreciate the sub-optimality of the uniform strategy, which wastes a portion of its budget in gathering samples that are significantly discounted, and, therefore, their weight in the estimator's variance shrinks to 0. On the other hand, the robust strategy and RIDO avoid this pitfall thanks to the exploitation of the discount factor, thus obtaining reduced variance estimates. Finally, we remark that RIDO has achieved the most competitive performance across various domains, values of the discount factor, and budget, thus clearly highlighting the benefits of adaptive strategies w.r.t. pre-determined ones.

6 Conclusions and Future Works

In this work, we studied the problem of allocating a budget Λ of transitions in the context of Monte Carlo policy evaluation to reduce the error of the policy expected return estimate. Leveraging the formalism of Data Collection Strategy (DCS) to model how an agent spends its interaction budget, we started by analyzing, in closed form, the variance of an unbiased return estimator for any possible DCS. Our result reveals that DCSs determined prior to the interaction with the environment (e.g., the usual uniform-in-the-horizon one and the robust one of Poiani et al. [2023]) fail to satisfy the ultimate goal of policy evaluation, i.e., produce a low error estimate. Furthermore, it also suggests that algorithms that spend the available budget Λ iteratively might successfully adapt their strategy to minimize the variance of the return estimator. Inspired by these findings, we propose an *adaptive*

[6]The uncertainty, in this case, arises both from the noise of the system together with the stochasticity of the initial state distribution.

method, RIDO, that, by exploiting information that has already been collected, can dynamically adapt its DCS to allocate a larger portion of transitions in time steps in which more accurate sampling is required to reduce the variance of the final estimate. After conducting a theoretical analysis on the properties of the proposed method, we present empirical studies that confirm its adaptivity across a different number of domains, values of budget Λ, and discount factors γ.

Our study offers exciting possibilities for future research. For example, although in this work we focused on the policy evaluation setting, it would be interesting to extend our ideas to policy search algorithms (e.g., Williams [1992]), with the goal of finding DCSs that minimize the variance of the empirical gradient that is adopted in the update rule. Furthermore, we notice that, since our approach is purely based on MC simulation, it does not fully leverage the Markovian properties of the underlying MDP. Combining TD techniques [Sutton and Barto, 2018] with mechanisms that truncate trajectories is a challenging and open research question that could lead to further improvements in the efficiency of RL algorithms.

References

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

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

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

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Karl W Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. Phasic policy gradient. In *International Conference on Machine Learning*, pages 2020–2027. PMLR, 2021.

Ruth F Curtain. Linear-quadratic control: An introduction, 1997.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *COLT*, volume 2, pages 255–270. Springer, 2002.

Amir-massoud Farahmand, Daniel Nikovski, Yuji Igarashi, and Hiroki Konaka. Truncated approximate dynamic programming with task-dependent terminal value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Josiah P Hanna, Philip S Thomas, Peter Stone, and Scott Niekum. Data-efficient policy evaluation through behavior policy search. In *International Conference on Machine Learning*, pages 1394–1403. PMLR, 2017.

Timothy Classen Hesterberg. *Advances in importance sampling*. Stanford University, 1988.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

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

Donghwan Lee and Niao He. Target-based temporal-difference learning. In *International Conference on Machine Learning*, pages 3713–3722. PMLR, 2019.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization, 2009.

Subhojyoti Mukherjee, Josiah P Hanna, and Robert D Nowak. Revar: Strengthening policy evaluation via reduced variance sampling. In *Uncertainty in Artificial Intelligence*, pages 1413–1422. PMLR, 2022.

Nhat M Nguyen, Abhineet Singh, and Kenneth Tran. Improving model-based rl with adaptive rollout using uncertainty estimation. 2018.

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

Brendan O'Donoghue, Ian Osband, Remi Munos, and Volodymyr Mnih. The uncertainty bellman equation and exploration. In *International Conference on Machine Learning*, pages 3836–3845, 2018.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients. *Machine Learning*, pages 1–57, 2022.

Riccardo Poiani, Alberto Maria Metelli, and Marcello Restelli. Multi-fidelity best-arm identification. *Advances in Neural Information Processing Systems*, 35:17857–17870, 2022.

Riccardo Poiani, Alberto Maria Metelli, and Marcello Restelli. Truncating trajectories in monte carlo reinforcement learning, 2023.

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

Chao Qu, Shie Mannor, and Huan Xu. Nonlinear distributional gradient temporal-difference learning. In *International Conference on Machine Learning*, pages 5251–5260. PMLR, 2019.

Antonin Raffin. Rl baselines3 zoo. `https://github.com/DLR-RM/rl-baselines3-zoo`, 2020.

Carlos Riquelme, Hugo Penedones, Damien Vincent, Hartmut Maennel, Sylvain Gelly, Timothy A Mann, Andre Barreto, and Gergely Neu. Adaptive temporal-difference learning for policy evaluation with per-state uncertainty estimates. *Advances in Neural Information Processing Systems*, 32, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Satinder P Singh and Richard S Sutton. Reinforcement learning with replacing eligibility traces. *Machine learning*, 22(1-3):123–158, 1996.

Wen Sun, J Andrew Bagnell, and Byron Boots. Truncated horizon policy search: Combining reinforcement learning & imitation learning. *arXiv preprint arXiv:1805.11240*, 2018.

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

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

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

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

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

Junjie Zhang, Jiafei Lyu, Xiaoteng Ma, Jiangpeng Yan, Jun Yang, Le Wan, and Xiu Li. Uncertainty-driven trajectory truncation for model-based offline reinforcement learning. *arXiv preprint arXiv:2304.04660*, 2023.

Rujie Zhong, Duohan Zhang, Lukas Schäfer, Stefano Albrecht, and Josiah Hanna. Robust on-policy sampling for data-efficient policy evaluation in reinforcement learning. *Advances in Neural Information Processing Systems*, 35:37376–37388, 2022.

A Related Works

Prior to delving into our theoretical analysis, we present a comprehensive analysis and discussion of previous works that are closely connected to our own research.

First of all, our work focuses on estimating a policy's performance in a given MDP [Sutton and Barto, 2018]. Considering the significance of this task, reducing the variance, or more generally, the error of the return estimator, is a problem that has received significant attention in the literature. A vast family of approaches that can be used to solve this problem deeply exploits the Markovian properties of the environment by relying on Temporal Difference [TD, see, e.g., Singh and Sutton, 1996, Sutton, 1988, Lee and He, 2019, Riquelme et al., 2019, Qu et al., 2019] learning. On the other hand, our work focuses purely on Monte Carlo simulation, which, differently from TD learning, can be transparently applied to non-Markovian environments. Another relevant line of work deals with optimizing the agent's policy to collect data within an environment (i.e., *behavioral* policy) to reduce the variance of an unbiased estimator for the return of a different *target* policy [Hanna et al., 2017, Zhong et al., 2022, Mukherjee et al., 2022]. These techniques are referred to as off-policy evaluation methods and usually rely on Importance Sampling [e.g, Hesterberg, 1988, Owen, 2013] techniques to guarantee the unbiasedness of the resulting estimate (i.e., to correct the distribution mismatch between the behavioral policy and the target one). However, we notice that these studies significantly differ from ours in that, instead of aiming for a behavior policy that reduces the estimator variance, we aim to directly exploit the properties of Monte Carlo data collection to reduce the on-policy estimator variance.

In the context of RL, exploration bonuses are widely adopted in control (where the goal is learning an optimal policy) to tackle the exploration-exploitation dilemma [e.g., Brafman and Tennenholtz, 2002, Auer et al., 2008, Tang et al., 2017, Jin et al., 2018, O'Donoghue et al., 2018, Zanette and Brunskill, 2019]. Initially, when the agent has limited knowledge about the environment, the exploration bonuses drive it to explore widely. As the agent's knowledge improves, the exploration bonuses decrease, and the agent can shift towards exploiting its learned policy more. In our work, instead, we use exploration bonuses to introduce a source of robustness w.r.t. the objective function that we are interested in, i.e., the variance of the return estimator for a given data collection strategy.

The work that is most related to ours is Poiani et al. [2023], where, the concept of truncating trajectories has been analyzed in the context of Monte Carlo RL. More specifically, the authors derived a *non-adaptive* schedule of trajectories that provably minimizes confidence intervals around the return estimator. In this work, on the other hand, we have shown the sub-optimality of pre-determined schedules, and we designed an *adaptive* algorithm that aims at minimizing the variance of the final estimate. The concept of truncating trajectories has also received some attention in other fields of research such as model-based policy optimization [Nguyen et al., 2018, Janner et al., 2019, Bhatia et al., 2022, Zhang et al., 2023], multi-task RL [Farahmand et al., 2016] and imitation learning [Sun et al., 2018]. However, in all these works, the motivation, the method, and the analysis completely differ w.r.t. what has been considered here. Finally, it is worth to mention that the concept of truncating trajectories in Monte Carlo RL drew inspiration from a recent work in the field of multi-fidelity bandit [Poiani et al., 2022], where the authors considered the idea of cutting trajectories while interacting with the environment to obtain a biased estimate of the return of a policy in planning algorithms such as depth-first search.

B Proofs and Derivations

In this section, we provide complete proofs of our theoretical results. More specifically, Section B.1 contains the proof of Theorem 3.1; Section B.2 the proof of Theorem 4.1, and Section B.3 proofs and details of additional statements that have been made in the main text (i.e., formal description of the transformation between optimization problems and how we applied this technique in RIDO, difficulties in deriving closed-form solutions for the optimization problems of interest, formal statement and proof of Equation (9), sub-optimality examples of non-adaptive methods whose variance cannot scale with the variance of the optimal DCS, as RIDO, instead, does).

B.1 Proof of Theorem 3.1

Theorem 3.1. *Consider a generic DCS $\boldsymbol{m}$ such that $m_T \geq 1$, then:*

$$\mathbb{V}\mathrm{ar}_{\boldsymbol{m}}\left[\hat{J}_{\boldsymbol{m}}(\pi)\right] = \sum_{t=0}^{T-1} \frac{1}{n_t}\left(\gamma^{2t}\mathbb{V}\mathrm{ar}(R_t) + 2\sum_{t'=t+1}^{T-1}\gamma^{t+t'}\mathbb{C}\mathrm{ov}(R_t, R_{t'})\right) =: \sum_{t=0}^{T-1}\frac{f_t}{n_t}. \tag{3}$$

Proof. Given that the different trajectories are independent, we have that:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_{\boldsymbol{m}}\left[\hat{J}_{\boldsymbol{m}}(\pi)\right] &= \sum_{h=1}^{T} m_h \mathbb{V}\mathrm{ar}\left[\sum_{t=0}^{h-1}\frac{\gamma^t R_t}{n_t}\right] \\
&= \sum_{h=1}^{T} m_h \sum_{t=0}^{h-1}\mathbb{V}\mathrm{ar}\left[\frac{\gamma^t R_t}{n_t}\right] + \sum_{h=1}^{T} m_h \sum_{t=0}^{h-2}\sum_{t'=t+1}^{h-1} 2\mathbb{C}\mathrm{ov}\left(\frac{\gamma^t R_t}{n_t}, \frac{\gamma^{t'} R_{t'}}{n_{t'}}\right) \\
&= \sum_{h=1}^{T} m_h \sum_{t=0}^{h-1}\frac{\gamma^{2t}}{n_t^2}\mathbb{V}\mathrm{ar}[R_t] + \sum_{h=1}^{T} m_h \sum_{t=0}^{h-2}\sum_{t'=t+1}^{h-1} 2\frac{\gamma^{t+t'}}{n_t n_{t'}}\mathbb{C}\mathrm{ov}\left(R_t, R_{t'}\right),
\end{aligned}$$

where the first step follows from the fact that different trajectories are independent, the second one from the variance of the sum of dependent random variable, namely $\mathbb{V}\mathrm{ar}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{V}\mathrm{ar}[X_i] + \sum_{i\neq j}\mathbb{C}\mathrm{ov}(X_i, X_j)$, and the third one by the fact that $\mathbb{V}\mathrm{ar}[aX] = a^2\mathbb{V}\mathrm{ar}[X]$ for some scalar $a \in \mathbb{R}$ and $\mathbb{C}\mathrm{ov}(aX, bY) = ab\mathbb{C}\mathrm{ov}(X, Y)$ for scalars $a, b \in \mathbb{R}$.

At this point, focus on:

$$\sum_{h=1}^{T} m_h \sum_{t=0}^{h-1}\frac{\gamma^{2t}}{n_t^2}\mathbb{V}\mathrm{ar}[R_t],$$

and fix $\bar{t} \in \{0, \dots, T-1\}$. By unrolling the summation, we notice that its contribution appears only in all h such that $h > \bar{t}$, thus leading to:

$$\sum_{h=1}^{T} m_h \sum_{t=0}^{h-1}\frac{\gamma^{2t}}{n_t^2}\mathbb{V}\mathrm{ar}[R_t], = \sum_{t=0}^{T-1}\frac{\gamma^{2t}}{n_t^2}\mathbb{V}\mathrm{ar}[R_t]\sum_{h=t+1}^{T} m_h.$$

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

$$\sum_{h=t+1}^{T} m_h = n_t - n_{t+1} + n_{t+1} - n_{t-2} + \dots + n_{T-2} - n_{T-1} + n_{T-1} = n_t. \tag{10}$$

Therefore:

$$\sum_{h=1}^{T} m_h \sum_{t=0}^{h-1}\frac{\gamma^{2t}}{n_t^2}\mathbb{V}\mathrm{ar}[R_t] = \sum_{t=0}^{T-1}\frac{\gamma^{2t}}{n_t}\mathbb{V}\mathrm{ar}[R_t]. \tag{11}$$

Now, let us focus on:

$$\sum_{h=1}^{T} m_h \sum_{t=0}^{h-2}\sum_{t'=t+1}^{h-1} 2\frac{\gamma^{t+t'}}{n_t n_{t'}}\mathbb{C}\mathrm{ov}\left(R_t, R_{t'}\right). \tag{12}$$

Fix the the index of the outer summation over time by considering $\bar{t} \in \{0, \dots, T-2\}$. By unrolling the summation, we notice that its contribution appears only in all h such that $h > \bar{t}+1$. For this reason, Equation (12) can be rewritten as:

$$\sum_{h=1}^{T} m_h \sum_{t=0}^{h-2}\sum_{t'=t+1}^{h-1} 2\frac{\gamma^{t+t'}}{n_t n_{t'}}\mathbb{C}\mathrm{ov}\left(R_t, R_{t'}\right) = \sum_{t=0}^{T-2}\sum_{h=t+2}^{T} m_h \sum_{t'=t+1}^{h-1}\frac{2\gamma^{t+t'}}{n_t n_{t'}}\mathbb{C}\mathrm{ov}\left(R_t, R_{t'}\right). \tag{13}$$

At this point, fix again $\bar{t} \in \{0, \dots, T-2\}$ as the index of the outer summation, and consider $t' \geq \bar{t}+1$. By unrolling the summation, we notice that t' appears only for $h > t'$. For this reason, we can rewrite Equation (13) as:

$$\sum_{t=0}^{T-2} \sum_{h=t+2}^{T} m_h \sum_{t'=t+1}^{h-1} \frac{2\gamma^{t+t'}}{n_t n_{t'}} \mathbb{C}\text{ov}\left(R_t, R_{t'}\right) = \sum_{t=0}^{T-2} \sum_{t'=t+1}^{T-1} \frac{2\gamma^{t+t'}}{n_t n_{t'}} \mathbb{C}\text{ov}\left(R_t, R_{t'}\right) \left(\sum_{h=t'+1}^{T} m_h \right). \tag{14}$$

However, by Equation (10), we obtain that:

$$\sum_{h=t'+1}^{T} m_h = n_{t'},$$

thus leading to:

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

Combining Equation (11) and (15) concludes the proof.

□

B.2 Proof of Theorem 4.1

To prove Theorem 4.1, we first provide some preliminaries lemmas on the properties of the optimization problems that we are considering, togheter with some technical results that will be used in our proofs. Then, we will move towards the analysis of RIDO.

B.2.1 Preliminaries for the proof of Theorem 4.1

We begin by proving the fact that for any timestep t in which f_t is negative, that there exists some future timestep t' such that $\sum_{i=t}^{t'} f_i \geq 0$.

Lemma B.1 (Variance function property)**.** *Consider* $f_t = \gamma^{2t}\mathbb{V}\text{ar}R_t + 2\sum_{t'=t+1}^{T-1} \gamma^{t+t'} \mathbb{C}\text{ov}(R_t, R_{t'})$. *For any* $t \in \{0, \dots, T-2\}$ *such that* $f_t < 0$*, there exists* $\bar{t} > t$ *such that* $\sum_{i=t}^{\bar{t}} f_t \geq 0$.

Proof. We proceed by contradiction. Suppose the claim to be false, then we would have:

$$\sum_{i=t}^{T-1} f_i < 0.$$

However, by manipulating $\sum_{i=t}^{T-1} f_i$, we obtain:

$$\sum_{i=t}^{T-1} f_i = \sum_{i=t}^{T-1} \gamma^{2i}\mathbb{V}\text{ar}R_i + 2\sum_{t'=i+1}^{T-2} \gamma^{i+t'} \mathbb{C}\text{ov}(R_i, R_{t'}) = \mathbb{V}\text{ar}\left[\sum_{i=t}^{T-1} \gamma^i R_i\right],$$

which is always greater or equal than 0, thus concluding the proof. □

We then continue by proving the result of Section 4.1 that justifies the transformation between optimization problems. However, rather than considering directly the optimization problem we are interested in (i.e., the one defined with f_t), we focus on a generalization that consider arbitrary vectors that satisfy the same properties as the one of Lemma B.1.

Lemma B.2 (Optimization of Variance-like functions)**.** *Let* $c = (c_1, \dots, c_k)$ *with* $c_i \in \mathbb{R}$*, such that* $c_1 < 0$*,* $\sum_{i=1}^{\bar{k}} c_i \leq 0$ *for all* $\bar{k} < k$*, and* $\sum_{i=1}^{k} c_i \geq 0$*. Let* $\Lambda \geq k$ *and consider the following*

optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \sum_{i=1}^{K} \frac{c_t}{x_t} \\ \textit{s.t.} \quad & \sum_{t=0}^{T-1} x_t = \Lambda \\ & x_i \geq x_{i+1}, \quad \forall i \in \{1, \dots, K-1\} \\ & x_i \geq 1, \quad \forall i \in \{1, \dots, K\}. \end{aligned} \tag{16}$$

Then, $\bar{\boldsymbol{x}} = \left(\frac{\Lambda}{k}, \dots, \frac{\Lambda}{k}\right)$ *is an optimal solution of* (16).

Proof. If $\bar{\boldsymbol{x}}$ is an optimal solution of (16), for all $\boldsymbol{x} = (x_1, \dots, x_k)$ that belongs to the feasible region it holds that:

$$\sum_{i=1}^{k} \frac{c_i}{\Lambda/k} \leq \sum_{i=1}^{k} \frac{c_i}{x_i},$$

which can be rewritten as:

$$\sum_{i=1}^{k-1} \frac{c_i}{\Lambda/k} \leq c_k \left(\frac{1}{x_k} - \frac{1}{\Lambda/k}\right) + \sum_{i=1}^{k-1} \frac{c_i}{x_i} = c_k \frac{\Lambda - k x_k}{x_k \Lambda} + \sum_{i=1}^{k-1} \frac{c_i}{x_i}. \tag{17}$$

At this point, we notice that $\Lambda \geq k x_k$ for any $\boldsymbol{x}$ that belongs to the feasible region. Furthermore, $\sum_{i=1}^{k} c_i \geq 0$, implies that $c_k \geq -\sum_{i=1}^{k-1} c_i \geq 0$. Therefore, a sufficient for Equation (17) to hold is that:

$$\sum_{i=1}^{k-1} \frac{c_i}{\Lambda/k} \leq \sum_{i=1}^{k-1} \frac{c_i}{x_i} - \sum_{i=1}^{k-1} c_i \left(\frac{1}{x_k} - \frac{1}{\Lambda/k}\right),$$

that can be rewritten as:

$$\sum_{i=1}^{k-1} c_i \left(\frac{1}{x_i} - \frac{1}{x_k}\right) \geq 0,$$

or equivalently:

$$\sum_{i=1}^{k-1} c_i \leq \sum_{i=1}^{k-1} c_i \frac{x_k}{x_i}. \tag{18}$$

However, as we shall show, Equation (18) is always satisfied. Indeed, since $\sum_{i=1}^{k-1} c_i \leq 0$ and $x_k \leq x_{k-1}$ we have that:

$$\sum_{i=1}^{k-1} c_i \leq \sum_{i=1}^{k-1} c_i \frac{x_k}{x_{k-1}} = c_{k-1} \frac{x_k}{x_{k-1}} + \sum_{i=1}^{k-2} c_i \frac{x_k}{x_{k-1}}.$$

Moreover, since $\sum_{i=1}^{k-1} c_i \leq 0$ and since $x_{k-1} \leq x_{k-2}$,

$$c_{k-1} \frac{x_k}{x_{k-1}} + \sum_{i=1}^{k-2} c_i \frac{x_k}{x_{k-1}} \leq c_{k-1} \frac{x_k}{x_{k-1}} + c_{k-2} \frac{x_k}{x_{k-2}} + \sum_{i=1}^{k-2} c_i \frac{x_k}{x_{k-2}}.$$

The properties that $x_i \geq x_{i+1}$ together with the fact that $\sum_{i=1}^{\bar{k}} c_i \leq 0$ for any $\bar{k} < k$ allows to iterate the process, thus concluding the proof. □

As one can see, applying multiple times Lemma B.2, to the problem we are considering, we obtain a transformed problem that is convex, since the objective function will be composed of summation of convex functions. We will provide additional details on this point later on. We now continue by studying the properties of optimization problems whose objective function satisfies the condition of Lemma B.2. More specifically, the following Lemma allows us to quantify the difference in the optimal solution when changing the budget constraint.

Lemma B.3 (Budget sensitivity analysis)**.** *Let $c_t \in \mathbb{R}$ for each $t \in \{0, \dots, T-1\}$. Define $\mathcal{Y} = \{i \in \{0, \dots, T-1\} : c_i < 0\}$. Let $y \in \mathcal{Y}$, and define $q(y)$ as the smallest integer in $\{y+1, \dots, T-1\}$ such that $\sum_{i=y}^{q(y)} c_i \geq 0$. Suppose that $q(y)$ is well-defined for any $y \in \mathcal{Y}$.*

Consider the following optimization problems:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \sum_{t=0}^{T-1} \frac{c_t}{x_t} \\ \textit{s.t.} \quad & \sum_{t=0}^{T-1} x_t = \Lambda \\ & x_t \geq x_{t+1}, \quad \forall t \in \{0, \dots, T-2\} \\ & x_t \geq 0, \quad \forall t \in \{0, \dots, T-1\} \\ & x_y = x_{y+1} = \dots = x_{q(y)}, \quad \forall y \in \mathcal{Y}, \end{aligned} \tag{19}$$

and,

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \sum_{t=0}^{T-1} \frac{c_t}{x_t} \\ \textit{s.t.} \quad & \sum_{t=0}^{T-1} x_t = \Lambda' \\ & x_t \geq x_{t+1}, \quad \forall t \in \{0, \dots, T-2\} \\ & x_t \geq 0, \quad \forall t \in \{0, \dots, T-1\} \\ & x_y = x_{y+1} = \dots = x_{q(y)}, \quad \forall y \in \mathcal{Y}, \end{aligned} \tag{20}$$

where $\Lambda, \Lambda' \in \mathbb{R}$ such that $\Lambda \geq T$ and $\Lambda' \geq T$. Define $\alpha = \frac{\Lambda'}{\Lambda}$ and consider $\boldsymbol{x}^*$ an optimal solution of (19)*. Then, $\alpha \boldsymbol{x}^*$ is an optimal solution of* (20)*.*

Proof. First of all, it is important to notice that both problems takes finite and positive value. This directly follow from the equality constraints, together with the fact that $q(y)$ is well-defined for any $y \in \mathcal{Y}$.

We now continue in proving the claim. Proceed by contradiction and suppose that $\alpha \boldsymbol{x}^*$ is not an optimal solution of (20), and let $\bar{\boldsymbol{x}}$ be an optimal solution of (20).

At this point, first of all, we notice that $\alpha \boldsymbol{x}^*$ is a feasible solution of (20). Indeed, we have that $\alpha x_t^* \geq 0$, $\alpha x_t^* \geq \alpha x_{t+1}^*$, $\alpha \sum_{t=0}^{T-1} x_t^* = \alpha \Lambda = \frac{\Lambda'}{\Lambda} \Lambda = \Lambda'$, and for all $y \in \mathcal{Y}$, $\alpha x_y^* = \alpha x_{y+1}^* = \dots = \alpha x_{q(y)}^*$.

Therefore, we can write:

$$\sum_{t=0}^{T-1} \frac{c_t}{\bar{x}_t^*} < \sum_{t=0}^{T-1} \frac{c_t}{\alpha x_t^*} = \frac{1}{\alpha} \sum_{t=0}^{T-1} \frac{c_t}{x_t^*}.$$

From which it follows that:

$$\sum_{t=0}^{T-1} \frac{c_t}{x_t^*} > \sum_{t=0}^{T-1} \frac{c_t}{\bar{x}_t^* / \alpha}.$$

However, for similar reasoning w.r.t. to the ones presented above, $\left(\bar{x}_1^*/\alpha, \dots, \bar{x}_{T-1}^*/\alpha\right)$ is a feasible solution for (19), from which it follows that $\boldsymbol{x}^*$ would not be optimal, which is impossible.

□

The following result, instead, is a technical Lemma that will be used to analyze the error that RIDO accumulates in each optimization round.

Lemma B.4 (Technical lemma)**.** *Consider a sequence of $K \in \mathbb{N}$ elements $(a_1, \dots, a_K)$ such that $a_i \in \mathbb{R}$ and $a_i > 0$ for all $i \in [K]$. Then:*

$$\frac{1}{\sum_{i=1}^{K} a_i} \leq \frac{1}{K^2} \sum_{i=1}^{K} \frac{1}{a_i}. \tag{21}$$

Proof. We begin with some notation. Consider $K \in \mathbb{N}$ such that $K > 1$, we denote with $\mathcal{V}_K$ the subset of entry-wise strictly positive vectors of $\mathbb{R}^K$, namely:

$$\mathcal{V}_K = \left\{ (a_1, \dots, a_K) \in \mathbb{R}^K | a_i > 0 \text{ for all } i \in [K] \right\}.$$

We now proceed by induction on K.

Consider $K = 1$ and $\boldsymbol{v} = (a_1) \in \mathcal{V}_1$. In this case, Equation (21) holds for all $\boldsymbol{v} \in \mathcal{V}_1$ since it reduces to:

$$\frac{1}{a_1} \leq \frac{1}{a_1}.$$

At this point, suppose that:

$$\frac{1}{\sum_{i=1}^{K} a_i} \leq \frac{1}{K^2} \sum_{i=1}^{K} \frac{1}{a_i},$$

holds for K and for all vectors $\boldsymbol{v}_K \in \mathcal{V}_K$, and consider:

$$\frac{1}{\sum_{i=1}^{K+1} a_i} \leq \frac{1}{(K+1)^2} \sum_{i=1}^{K+1} \frac{1}{a_i},$$

for any vector $\boldsymbol{v}_{K+1} = (a_1, \dots a_{K+1}) \in \mathcal{V}_{K+1}$. At this point, notice that, for all $\boldsymbol{v}_{K+1} \in \mathcal{V}_{K+1}$ the vector $\boldsymbol{v}_{K,-i}$ that is obtained from $\boldsymbol{v}_{K+1}$ by removing the i-th component belongs to $\mathcal{V}_K$. At this point, focus on:

$$\frac{1}{(K+1)^2} \sum_{i=1}^{K+1} \frac{1}{a_i} = \frac{1}{(K+1)^2} \left(\sum_{i=1}^{K} \frac{1}{a_i} + \frac{1}{a_{k+1}} \right).$$

Thanks to the inductive hypothesis and some algebraic manipulations, we have that:

$$\begin{aligned}
\frac{1}{(K+1)^2} \left(\sum_{i=1}^{K} \frac{1}{a_i} + \frac{1}{a_{k+1}} \right) &= \frac{K^2}{K^2(K+1)^2} \sum_{i=1}^{K} \frac{1}{a_i} + \frac{1}{(K+1)^2} \frac{1}{a_{k+1}} \\
&\geq \frac{K^2}{(K+1)^2} \left(\frac{1}{\sum_{i=1}^{K} a_i} \right) + \frac{1}{(K+1)^2 a_{K+1}} \\
&= \frac{K^2}{(K+1)^2} \left(\frac{1}{\sum_{i=1}^{K} a_i} + \frac{1}{a_{K+1} K^2} \right).
\end{aligned}$$

At this point, we need to show that:

$$\frac{K^2}{(K+1)^2} \left(\frac{1}{\sum_{i=1}^{K} a_i} + \frac{1}{a_{K+1} K^2} \right) \geq \frac{1}{\sum_{i=1}^{K} a_i + a_{K+1}},$$

holds. Set, for the sake of exposition $c = \sum_{i=1}^{K} a_i$ and $d = a_{K+1}$. Then, we can rewrite the previous inequality as:

$$\frac{K^2}{(K+1)^2} \left(\frac{1}{c} + \frac{1}{K^2 d} \right) \geq \frac{1}{c+d}.$$

Rearranging the terms we obtain:

$$\frac{K^2}{(K+1)^2} \left(\frac{K^2 d + c}{c d K^2} \right) \geq \frac{1}{c+d}.$$

Which, in turns, lead to:

$$K^2(K^2d + c)(c + d) \geq (K + 1)^2 cdK^2.$$

Multiplying each term and dividing by K^2 leads to:

$$d^2K^2 - cdK + c^2 \geq 0,$$

which holds for any value of $K > 0$, and $d, c > 0$, thus concluding the proof. □

Finally, the following Lemma will be used to take into account the rounding effect that comes from solving a continuous relaxation rather than an integer optimization problem.

Lemma B.5 (Rounding effect error)**.** *Consider a generic T-dimensional vector $\boldsymbol{n} = (n_0, \dots n_{T-1})$ such that $n_i \geq 1$ for all $i \in \{0, \dots, T-1\}$. Let $q = \sum_{t=0}^{T-1} n_t$, and define $k = q - \sum_{t=0}^{T-1} \lfloor n_t \rfloor$. Consider the vector $\bar{\boldsymbol{n}} = (\bar{n}_0, \dots \bar{n}_{T-1})$ such that:*

$$\bar{n}_t = \lfloor n_t \rfloor + \mathbf{1}\{t < k\}.$$

Define $g(\boldsymbol{n}) = \sum_{t=0}^{T-1} \frac{c_t}{n_t}$ for some vector $\mathbf{c} = (c_0, \dots, c_{T-1})$ with $c_t \in \mathbb{R}$. Then, the following holds:

$$\sum_{t:c_t \geq 0} \frac{c_t}{\bar{n}_t} \leq 2 \sum_{t:c_t \geq 0} \frac{c_t}{n_t}, \tag{22}$$

$$\sum_{t:c_t \leq 0} \frac{c_t}{\bar{n}_t} \leq \frac{1}{2} \sum_{t:c_t \leq 0} \frac{c_t}{n_t}. \tag{23}$$

Proof. We begin by proving Equation (22). First of all, let us notice that:

$$\sum_{t:c_t \geq 0} \frac{c_t}{n_t} \geq \sum_{t:c_t \geq 0} \frac{c_t}{\bar{n}_t + 1} \geq \sum_{t:c_t \geq 0} \frac{c_t}{2\bar{n}_t}, \tag{24}$$

where in the first inequality we have used $c_t \geq 0$ together with $|n_t - \bar{n}_t| \leq 1$, while in the second one we have used $c_t \geq 0$ together with $\bar{n}_t \geq 1$. Equation (22) directly follows from Equation (24).

We continue by proving Equation (23). Similar to Equation (24), it is possible to obtain:

$$\sum_{t:c_t \leq 0} \frac{c_t}{\bar{n}_t} \leq \sum_{t:c_t \leq 0} \frac{c_t}{n_t + 1} \leq \sum_{t:c_t \leq 0} \frac{c_t}{2n_t} = \frac{1}{2} \sum_{t:c_t \leq 0} \frac{c_t}{n_t}, \tag{25}$$

where in the first step we have used $c_t \leq 0$ together with $c_t \leq 0$, while in the second one we have used $c_t \leq 0$ together with $n_t \geq 1$. □

B.2.2 RIDO analysis

We begin with some concentration inequalities. We report for completeness the result (Theorem 10) of Maurer and Pontil [2009] that we use to construct confidence intervals around the standard deviation.

Lemma B.6 (Standard deviation confidence intervals)**.** *Let $n \geq 2$ and consider $X_1, \dots, X_n$ be i.i.d. random variables with values in $[0, 1]$. Define:*

$$\hat{\sigma} = \sqrt{\frac{1}{n(n-1)} \sum_{i<j} (X_i - X_j)^2}.$$

Then, for $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have that:

$$|\hat{\sigma} - \sigma| \leq \sqrt{\frac{2\ln(1/\delta)}{n-1}},$$

where $\sigma = \mathbb{E}\hat{\sigma}$.

We then continue with similar results for the estimation of the covariances between random variables.

Lemma B.7 (Covariance confidence intervals)**.** *Consider $(X_1, Y_1), \dots (X_n, Y_n)$ i.i.d. random variables with values in $[0, 1]$ sampled from the joint distribution $f_{X,Y}$. Moreover, let $X_{n+1}, \dots, X_{n+k}$ be k i.i.d. random variables with values in $[0, 1]$ sampled from distribution $f_X = \mathbb{E}_Y [f_{X,Y}]$. Define, for all $i \in [n]$, $Z_i = X_i Y_i$, and let $\hat{z} = \frac{1}{n} \sum_{i=1}^{n} Z_i$, $\hat{x} = \frac{1}{n+k} \sum_{i=1}^{n+k} X_i$ and $\hat{y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$. Then, for $\delta \in (0, 1)$, we have that:*

$$|\mathbb{E}\hat{z} - \mathbb{E}\hat{x}\mathbb{E}\hat{y} - (\hat{z} - \hat{x}\hat{y})| \leq 3\sqrt{\frac{2\log(6/\delta)}{n}}.$$

Proof. By Hoeffding Inequality Boucheron et al. [2003], we have that, for some confidence level δ', the following holds with probability at least $1 - \delta'$:

$$|\hat{z} - \mathbb{E}\hat{z}| \leq \sqrt{\frac{2\log(2/\delta')}{n}},$$

and, similarly for $\hat{x}$ and $\hat{y}$. Therefore, by Boole's inequality, it follows that, with probability at least $1 - \delta$, we have that:

$$|\hat{z} - \mathbb{E}\hat{z}| \leq \sqrt{\frac{2\log(6/\delta)}{n}}, \tag{26}$$

and, similarly, for $\hat{x}$ and $\hat{y}$. [7]

Therefore, with probability at least $1 - \delta$ we have that:

$$\begin{aligned}
|\mathbb{E}\hat{z} - \mathbb{E}\hat{x}\mathbb{E}\hat{y} - (\hat{z} - \hat{x}\hat{y})| &\leq |\mathbb{E}\hat{z} - \hat{z}| + |\mathbb{E}\hat{x}\mathbb{E}\hat{y} - \hat{x}\hat{y}| \\
&\leq \sqrt{\frac{2\log(6/\delta)}{n}} + |\mathbb{E}\hat{x}\mathbb{E}\hat{y} - \hat{y}\mathbb{E}\hat{x} + \hat{y}\mathbb{E}x - \hat{x}\hat{y}| \\
&\leq \sqrt{\frac{2\log(6/\delta)}{n}} + |\mathbb{E}\hat{x}(\mathbb{E}\hat{y} - \hat{y})| + |\hat{y}(\mathbb{E}\hat{x} - \hat{x})| \\
&\leq 2\sqrt{\frac{2\log(6/\delta)}{n}} + |\hat{y}|\sqrt{\frac{2\log(6/\delta)}{n}} \\
&\leq 3\sqrt{\frac{2\log(6/\delta)}{n}}.
\end{aligned}$$

where we combined Equation (26) together with triangular inequalities. □

At this point, before diving into the presentation of the good event under which we will conduct our analysis, we provide a formal definition of our estimators. Consider a generic dataset of trajectories of different lenght. Define, for each $t \in \{0, \dots, T-1\}$:

$$\sqrt{\widehat{\mathbb{V}}\mathrm{ar}\left(R_t\right)} = \sqrt{\frac{1}{n_t(n_t - 1)} \sum_{1 \leq i < j \leq n} \left(R_t^{(i)} - R_t^{(j)}\right)^2}, \tag{27}$$

where $R_t^{(i)}$ denotes the reward gathered at step t in some trajectory whose length is at least $t + 1$. Moreover, for t, t' such that $t < t'$, define:

$$\widehat{\mathbb{C}}\mathrm{ov}(R_t, R_t') = \frac{1}{n_{t'}} \sum_{i=1}^{n_{t'}} R_t^{(i)} R_{t'}^{(i)} - \left(\frac{1}{n_t} \sum_{i=1}^{n_t} R_t^{(i)}\right)\left(\frac{1}{n_{t'}} \sum_{i=1}^{n_{t'}} R_{t'}^{(i)}\right). \tag{28}$$

Lemma B.8 (Good event)**.** *The following conditions holds for all phases of RIDO, with probability at least $1 - \delta$:*

$$\left|\sqrt{\mathbb{V}\mathrm{ar}\left(R_t\right)} - \sqrt{\widehat{\mathbb{V}}\mathrm{ar}_i\left(R_t\right)}\right| \leq \sqrt{\frac{2\log\left(\frac{6(T+T^2)\Lambda K}{\delta}\right)}{n_t}} = \boldsymbol{C}_{i,t}^{\sigma}. \tag{29}$$

and:

$$\left|\mathbb{C}\mathrm{ov}\left(R_t, R_{t'}\right) - \widehat{\mathbb{C}}\mathrm{ov}_i\left(R_t, R_{t'}\right)\right| \leq 3\sqrt{\frac{2\log\left(\frac{6(T+T^2)\Lambda K}{\delta}\right)}{n_{t'}}} = \boldsymbol{C}_{i,t,t'}^{c}. \tag{30}$$

[7] For $\hat{x}$ the confidence intervals holds with $\sqrt{\frac{2\log(6/\delta)}{n+k}}$, which is possibly smaller since $n \leq n + k$.

Proof. The proof follows by combining Lemma B.7 and Lemma B.6, and by taking the union bound over the different time steps, optimization rounds, and possible ways in which the budget can be spent. □

At this point, first we show that, with high probability, the objective function of the empirical optimization problem (5) satisfies the same property of the objective function of the original optimization problem (4), i.e., Lemma B.1. Consequently, it holds that the procedure described in the main text in Section 4.1 leads to a transformed convex optimization problem that preserves the optimal solution. For this reason, in the rest of this section, under the good event of Lemma B.8, we assume that RIDO has actually access to an optimal solution of the continuous relaxation of (5), which can be obtained in a computational efficient way by transforming the optimization problem.

Lemma B.9 (High probability property of the empirical problem)**.** *Let $\beta = \frac{6(T+T^2)\Lambda K}{\delta}$ and consider a generic phase i of Algorithm 1. Define:*

$$\hat{f}_{t,i} = \gamma^{2t}\left(\sqrt{\widehat{\mathbb{V}}\mathrm{ar}_i\left(R_t\right)} + \boldsymbol{C}^{\sigma}_{i,t}\right)^2 + 2\sum_{t'=t+1}^{T-1}\gamma^{t+t'}\left(\widehat{\mathbb{C}}\mathrm{ov}_i(R_t, R_{t'}) + \boldsymbol{C}^{c}_{i,t,t'}\right).$$

Suppose that $\hat{f}_{t,i} < 0$. Then, with probability at least $1-\delta$, for any $t \in \{0, \dots, T-2\}$ there exists $\bar{t} > t$ such that $\sum_{j=t}^{\bar{t}} \hat{f}_{j,i} \geq 0$ holds.

Proof. We proceed by contradiction. Suppose that $\hat{f}_{t,i} < 0$ and $\sum_{j=t}^{\bar{t}} \hat{f}_{j,i} < 0$ for all $\bar{t} > t$, and, thus, also for $\bar{t} = T-1$. Due to Lemma B.8, we have that:

$$\sum_{j=t}^{T-1}\hat{f}_{j,i} \geq \sum_{j=t}^{T-1}\gamma^{2j}\mathbb{V}\mathrm{ar}\left(R_j\right) + 2\sum_{t'=j+1}^{T-1}\gamma^{i+t'}\mathbb{C}\mathrm{ov}\left(R_j, R_{t'}\right) = \mathbb{V}\mathrm{ar}\left(\sum_{j=t}^{T-1}\gamma^j R_j\right),$$

which, however, is always greater or equal than 0, thus leading to a contradiction and concluding the proof. □

To analyze the performance of Algorithm 1, we will study the following quantity:

$$\mathbb{V}\mathrm{ar}_{\hat{\boldsymbol{n}}}\left[\hat{J}_{\hat{\boldsymbol{n}}}(\pi)\right] - \mathbb{V}\mathrm{ar}_{\boldsymbol{n}^*}\left[\hat{J}_{\boldsymbol{n}^*}(\pi)\right] = \sum_{t=0}^{T-1}\frac{f_t}{\sum_{i=0}^{K-1}\hat{n}_{t,i}} - \sum_{t=0}^{T-1}\frac{f_t}{n^*_t}. \tag{31}$$

More specifically, by upper bounding Equation (31) we implicitly upper-bound also the variance of the DCS computed by RIDO. At this point, we proceed by analyzing this quantity. The first step, which is presented in the following Lemma, stands in deriving a first errro decomposition on Equation (31).

Lemma B.10 (Error decomposition)**.** *Let $f_t = \gamma^{2t}\mathbb{V}\mathrm{ar}\left(R_t\right) + 2\sum_{t'=t+1}^{T-1}\gamma^{t+t'}\mathbb{C}\mathrm{ov}\left(R_t, R_{t'}\right)$. Let $y \in \mathcal{Y}$, and define $q(y)$ as the smallest integer in $\{y+1, \dots, T-1\}$ such that $\sum_{i=y}^{q(y)} f_i \geq 0$. Equation* (31) *can be upper bounded by:*

$$\begin{aligned}\frac{2}{K^2}\sum_{i=1}^{K}\sum_{t=0}^{T-1} f_t\left(\frac{1}{\bar{n}_{t,i}} - \frac{1}{\tilde{x}^*_t + 1}\right) + \frac{2}{K^2}\sum_{i=1}^{K}\sum_{t:f_t<0}\frac{|f_t|}{\bar{n}_{t,i}} - \sum_{t:f_t<0}\frac{|f_t|}{\sum_{i=1}^{k}\hat{n}_{t,i}} \\ + \frac{2}{K}\sum_{t=0}^{T-1}\frac{f_t}{\tilde{x}^*_t + 1} - \sum_{t=0}^{T-1}\frac{f_t}{n^*_t}.\end{aligned} \tag{32}$$

where $\bar{\boldsymbol{n}}_i$ is the optimal solution of the continuous relaxation (5)*, $\hat{\boldsymbol{n}}_i$ is the rounding DCS obtained from $\bar{\boldsymbol{n}}_i$, and $\tilde{x}^*$ is the optimal solution of the following optimization problem:*

$$\begin{aligned}
\min_{\boldsymbol{x}} \quad & \sum_{i=1}^{K} \frac{f_t}{x_t} \\
\textit{s.t.} \quad & \sum_{t=0}^{T-1} x_t = b - T \\
& x_i \geq x_{i+1}, \quad \forall i \in \{1, \dots, K-1\} \\
& x_i \geq 0, \quad \forall i \in \{1, \dots, K\} \\
& x_y = x_{y+1} = \dots = x_{q(y)}, \quad \forall y \in \mathcal{Y}.
\end{aligned} \tag{33}$$

Proof. Let us start by analyzing Equation (31):

$$\mathcal{R} = \sum_{t=0}^{T-1} \frac{f_t}{\sum_{i=1}^{k} \hat{n}_{t,i}} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*} = \sum_{t:f_t \geq 0} \frac{f_t}{\sum_{i=1}^{k} \hat{n}_{t,i}} + \sum_{t:f_t < 0} \frac{f_t}{\sum_{i=1}^{k} \hat{n}_{t,i}} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*}.$$

Due to Lemma B.4, we can upper the previous Equation obtaining:

$$\begin{aligned}
\mathcal{R} &\leq \frac{1}{K^2} \sum_{i=1}^{K} \sum_{t:f_t \geq 0} \frac{f_t}{\hat{n}_{t,i}} + \sum_{t:f_t < 0} \frac{f_t}{\sum_{i=1}^{k} \hat{n}_{t,i}} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*} \\
&\leq \frac{2}{K^2} \sum_{i=1}^{K} \sum_{t:f_t \geq 0} \frac{f_t}{\bar{n}_{t,i}} + \sum_{t:f_t < 0} \frac{f_t}{\sum_{i=1}^{k} \hat{n}_{t,i}} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*} \\
&= \frac{2}{K^2} \sum_{i=1}^{K} \sum_{t=0}^{T-1} \frac{f_t}{\bar{n}_{t,i}} - \frac{2}{K^2} \sum_{i=1}^{K} \sum_{t:f_t < 0} \frac{f_t}{\bar{n}_{t,i}} + \sum_{t:f_t < 0} \frac{f_t}{\sum_{i=1}^{k} \hat{n}_{t,i}} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*}.
\end{aligned}$$

where in the first step, we have used Lemma B.4, in the second one Lemma B.5, and in the third one we have added and subtracted $\frac{2}{K^2} \sum_{i=1}^{K} \sum_{t:f_t<0} \frac{f_t}{\bar{n}_{t,i}}$. The proof directly follows by adding and subtracting:

$$\frac{2}{K^2} \sum_{i=1}^{K} \sum_{t=0}^{T-1} \frac{f_t}{\tilde{x}_t^* + 1} = \frac{2}{K} \sum_{t=0}^{T-1} \frac{f_t}{\tilde{x}_t^* + 1}.$$

□

At this point, the following Lemma provides an upper bound on Equation (32). More specifically, we focus on the first term, that is:

$$\frac{2}{K^2} \sum_{i=1}^{K} \sum_{t=0}^{T-1} f_t \left(\frac{1}{\bar{n}_{t,i}} - \frac{1}{\tilde{x}_t^* + 1} \right),$$

which can be interpreted as the error that RIDO cumulates in its rounds.

Lemma B.11 (Cumulative error)**.** *Let $\beta = \frac{6(T+T^2)\Lambda K}{\delta}$. Let $y \in \mathcal{Y}$, and define $q(y)$ as the smallest integer in $\{y+1, \dots, T-1\}$ such that $\sum_{i=y}^{q(y)} f_i \geq 0$. Let $\tilde{\boldsymbol{x}}^*$ be the solution of the following optimization problem:*

$$\begin{aligned}
\min_{\boldsymbol{x}} \quad & \sum_{i=1}^{K} \frac{f_t}{x_t} \\
\textit{s.t.} \quad & \sum_{t=0}^{T-1} x_t = b - T \\
& x_i \geq x_{i+1}, \quad \forall i \in \{1, \dots, K-1\} \\
& x_i \geq 0, \quad \forall i \in \{1, \dots, K\} \\
& x_y = x_{y+1} = \dots = x_{q(y)}, \quad \forall y \in \mathcal{Y},
\end{aligned} \tag{34}$$

and let $\bar{\boldsymbol{n}}_i$ be the solution of the continuous relaxation of (5) *during phase i. Then, with probability at least $1-\delta$, the following holds:*

$$\frac{2}{K^2}\sum_{i=1}^{K}\sum_{t=0}^{T-1} f_t\left(\frac{1}{\bar{n}_{t,i}}-\frac{1}{\tilde{x}_t^*+1}\right) \leq \frac{192}{K^{\frac{3}{2}}}\log\left(\frac{6(T+T^2)\Lambda K}{\delta}\right)\left(\sum_{t=0}^{T-1}\gamma^t\right)^2 \tag{35}$$

Proof. The proof is split into 3 parts. In particular, we will analyze:

$$\frac{2}{K^2}\sum_{i=1}^{K}\sum_{t=0}^{T-1} f_t\left(\frac{1}{\bar{n}_{t,i}}-\frac{1}{\tilde{x}_t^*+1}\right)$$

first for a generic phase $i>1$, then for $i=1$, and finally we will put everything together.

Let us start by considering a generic phase $i>1$, and focus on:

$$\sum_{t=0}^{T-1} f_t\left(\frac{1}{\bar{n}_{t,i}}-\frac{1}{\tilde{x}_t^*+1}\right) \tag{36}$$

First of all, focus on $\sum_{t=0}^{T-1}\frac{f_t}{\tilde{x}_t+1}$. Let us define $\tilde{\boldsymbol{g}}=(\tilde{g}_0,\dots,\tilde{g}_{T-1})$ as the solution to the following optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{g}} \quad & \sum_{t=0}^{T-1}\frac{f_t}{g_t} \\
\text{s.t.} \quad & \sum_{t=0}^{T-1} g_t = b \\
& g_t \geq g_{t+1}, \quad \forall t\in\{0,\dots,T-2\} \\
& g_t \geq 1, \quad \forall t\in\{0,\dots,T-1\}.
\end{aligned} \tag{37}$$

It is easy to see that: [8]

$$\sum_{t=0}^{T-1}\frac{f_t}{\tilde{g}_t} \leq \sum_{t=0}^{T-1}\frac{f_t}{\tilde{x}_t^*+1}.$$

Plugging this result into Equation (36) leads to:

$$\sum_{t=0}^{T-1} f_t\left(\frac{1}{\bar{n}_{t,i}}-\frac{1}{\tilde{x}_t^*+1}\right) \leq \sum_{t=0}^{T-1} f_t\left(\frac{1}{\bar{n}_{t,i}}-\frac{1}{\tilde{g}_t^*}\right). \tag{38}$$

Due to Lemma B.8, with probability at least $1-\delta$, we can further upper bound Equation (38) with:

$$\sum_{t=0}^{T-1}\frac{\gamma^{2t}\left(\sqrt{\widehat{\mathbb{V}}\mathrm{ar}_i\left(R_t\right)}+\mathrm{C}_{i,t}^{\sigma}\right)^2}{\bar{n}_{t,i}} + 2\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\frac{\gamma^{t+t'}\left(\widehat{\mathbb{C}}\mathrm{ov}(R_t,R_{t'})+\mathrm{C}_{i,t,t'}^{c}\right)}{\bar{n}_{t,i}} - \sum_{t=0}^{T-1}\frac{f_t}{\tilde{g}_t},$$

However, since $\tilde{\boldsymbol{g}}$ is a feasible solution of the continuous relaxation of (5), and since $\bar{n}_{t,i}$ is the minimizer of the continuous relaxation of (5) at phase i, we can further bound the previous equation with:

$$\sum_{t=0}^{T-1}\frac{\gamma^{2t}\left(\sqrt{\widehat{\mathbb{V}}\mathrm{ar}_i\left(R_t\right)}+\mathrm{C}_{i,t}^{\sigma}\right)^2}{\tilde{g}_t} + 2\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\frac{\gamma^{t+t'}\left(\widehat{\mathbb{C}}\mathrm{ov}(R_t,R_{t'})+\mathrm{C}_{i,t,t'}^{c}\right)}{\tilde{g}_t} - \sum_{t=0}^{T-1}\frac{f_t}{\tilde{g}_t}.$$

[8] This step follows by considering the optimization problem that defines $\tilde{\boldsymbol{g}}$. With a change of variable $g_t = x_t+1$, we can notice that $\tilde{x}_t^*+1$ is indeed a feasible solution of the same optimization problem. Furthermore, notice that due to Lemma B.2, we can neglect the constraints on y.

Moreover, due to Lemma B.8, we can further upper-bound the previous Equation with:

$$\sum_{t=0}^{T-1} \frac{\gamma^{2t}\left(\sqrt{\mathbb{V}\mathrm{ar}\left[R_t\right]}+2\mathrm{C}_{i,t}^{\sigma}\right)^2}{\tilde{g}_t} + 2\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1} \frac{\gamma^{t+t'}\left(\mathbb{C}\mathrm{ov}(R_t,R_{t'})+2\mathrm{C}_{i,t,t'}^{c}\right)}{\tilde{g}_t} - \sum_{t=0}^{T-1}\frac{f_t}{\tilde{g}_t},$$

Let us now focus on:

$$\sum_{t=0}^{T-1} \frac{\gamma^{2t}\left(\sqrt{\mathbb{V}\mathrm{ar}\left[R_t\right]}+2\mathrm{C}_{i,t}^{\sigma}\right)^2}{\tilde{g}_t} + 2\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1} \frac{\gamma^{t+t'}\left(\mathbb{C}\mathrm{ov}(R_t,R_{t'})+2\mathrm{C}_{i,t,t'}^{c}\right)}{\tilde{g}_t}. \tag{39}$$

Equation (39) can be decomposed into:

$$\sum_{t=0}^{T-1} \frac{\gamma^{2t}4\sqrt{\mathbb{V}\mathrm{ar}\left[R_t\right]}\mathrm{C}_{i,t}^{\sigma}+4\gamma^{2t}\left(\mathrm{C}_{i,t}^{\sigma}\right)^2}{\tilde{g}_t} + 4\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1} \frac{\gamma^{t+t'}\mathrm{C}_{i,t,t'}^{c}}{\tilde{g}_t} \tag{40}$$

and,

$$\sum_{t=0}^{T-1}\frac{f_t}{\tilde{g}_t}. \tag{41}$$

Thus leading to:

$$\mathcal{R} \leq \sum_{t=0}^{T-1} \frac{\gamma^{2t}4\sqrt{\mathbb{V}\mathrm{ar}\left[R_t\right]}\mathrm{C}_{i,t}^{\sigma}+4\gamma^{2t}\left(\mathrm{C}_{i,t}^{\sigma}\right)^2}{\tilde{g}_t} + 4\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1} \frac{\gamma^{t+t'}\mathrm{C}_{i,t,t'}^{c}}{\tilde{g}_t} \tag{42}$$

We now proceed by bounding each term in Equation (42). Define, for brevity $h_{i,t} = \sum_{j=0}^{i-1}\hat{n}_{t,j}$. Let us first focus on:

$$\begin{aligned}
\sum_{t=0}^{T-1}\frac{4\gamma^{2t}\sqrt{\mathbb{V}\mathrm{ar}\left[R_t\right]}\mathrm{C}_{i,t}^{\sigma}}{\tilde{g}_t} &\leq \sum_{t=0}^{T-1}\frac{4\gamma^{2t}}{\tilde{g}_t}\sqrt{\frac{2\log\left(\frac{2(T+T^2)\Lambda K}{\delta}\right)}{h_{i-1,t}}} \\
&\leq 8\sqrt{\log\left(\frac{2(T+T^2)\Lambda K}{\delta}\right)}\sum_{t=0}^{T-1}\frac{\gamma^{2t}}{\tilde{g}_t\sqrt{i-1}} \\
&\leq 16\sqrt{\log\left(\frac{2(T+T^2)\Lambda K}{\delta}\right)}\sum_{t=0}^{T-1}\frac{\gamma^{2t}}{\sqrt{i}}.
\end{aligned}$$

where the first step follows from the definition of the confidence intervals, together with the fact that rewards are bounded in $[0,1]$, the second one by recalling that $h_{i-1,t} = \sum_{j=1}^{i-1}\hat{n}_{t,j} \geq i-1$, and the third one by noticing that $\sqrt{i} \leq 2\sqrt{i-1}$.

Similary, for what concerns:

$$\begin{aligned}
\sum_{t=0}^{T-1}\frac{4\gamma^{2t}\left(\mathrm{C}_{i,t}^{\sigma}\right)^2}{\tilde{g}_t} &\leq \sum_{t=0}^{T-1}\frac{4\gamma^{2t}}{\tilde{g}_t}\frac{2\log\left(\frac{2(T+T^2)\Lambda K}{\delta}\right)}{h_{i-1,t}} \\
&\leq 8\log\left(\frac{2(T+T^2)\Lambda K}{\delta}\right)\sum_{t=0}^{T-1}\frac{\gamma^{2t}}{\tilde{g}_t(i-1)} \\
&\leq 16\log\left(\frac{2(T+T^2)\Lambda K}{\delta}\right)\sum_{t=0}^{T-1}\frac{\gamma^{2t}}{\sqrt{i}}
\end{aligned}$$

Finally, what is left is:

$$\begin{aligned}
4\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\frac{\gamma^{t+t'}\mathrm{C}_{i,t,t'}^{c}}{\tilde{g}_t} &\leq 24\sqrt{\log\left(\frac{6(T+T^2)\Lambda K}{\delta}\right)}\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\frac{\gamma^{t+t'}}{\tilde{g}_t\sqrt{i-1}} \\
&\leq 48\sqrt{\log\left(\frac{6(T+T^2)\Lambda K}{\delta}\right)}\sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\frac{\gamma^{t+t'}}{\sqrt{i}}
\end{aligned}$$

For what concerns phase $i = 1$, instead, the budget is allocated uniformly. Therefore, we have that:

$$\sum_{t=0}^{T-1} f_t \left(\frac{1}{b/T} - \frac{1}{\tilde{g}_t} \right) \leq \sum_{t=0}^{T-1} \frac{f_t}{b/T} \leq \sum_{t=0}^{T-1} f_t \leq \left(\sum_{t=0}^{T-1} \gamma^t \right)^2$$

At this point, plugging these results into Equation (38) leads to:

$$\frac{2}{K^2} \sum_{i=1}^{K} \left(48 \log \left(\frac{6(T+T^2)\Lambda K}{\delta} \right) \left(\sum_{t=0}^{T-1} \gamma^t \right)^2 \right) \frac{1}{\sqrt{i}} \tag{43}$$

To conclude the proof, we notice that $\sum_{i=1}^{n} \frac{1}{\sqrt{i}} \leq 2\sqrt{n} - 1$, thus leading to:

$$\frac{192}{K^{\frac{3}{2}}} \log \left(\frac{6(T+T^2)\Lambda K}{\delta} \right) \left(\sum_{t=0}^{T-1} \gamma^t \right)^2$$

which is the desired result. □

We now continue by upper bounding another term of Equation (31), that is:

$$\frac{2}{K} \sum_{t=0}^{T-1} \frac{f_t}{\tilde{x}_t^* + 1} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*}.$$

Lemma B.12 (Exploration error)**.** *Let $y \in \mathcal{Y}$, and define $q(y)$ as the smallest integer in $\{y+1, \dots, T-1\}$ such that $\sum_{i=y}^{q(y)} f_i \geq 0$. Let $\tilde{\boldsymbol{x}}^*$ be the solution of the following optimization problem:*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \sum_{i=1}^{K} \frac{f_t}{x_t} \\ \textit{s.t.} \quad & \sum_{t=0}^{T-1} x_t = b - T \\ & x_i \geq x_{i+1}, \quad \forall i \in \{1, \dots, K-1\} \\ & x_i \geq 0, \quad \forall i \in \{1, \dots, K\} \\ & x_y = x_{y+1} = \dots = x_{q(y)}, \quad \forall y \in \mathcal{Y}. \end{aligned} \tag{44}$$

Then,

$$\frac{2}{K} \sum_{t=0}^{T-1} \frac{f_t}{\tilde{x}_t^* + 1} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*} \leq \frac{c+1}{c-1} \sum_{t=0}^{T-1} \frac{f_t}{x_t^*},$$

where c is such that $cT = b$, and $\boldsymbol{x}^*$ is the solution of the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \sum_{i=1}^{K} \frac{f_t}{x_t} \\ \textit{s.t.} \quad & \sum_{t=0}^{T-1} x_t = \Lambda \\ & x_i \geq x_{i+1}, \quad \forall i \in \{1, \dots, K-1\} \\ & x_i \geq 0, \quad \forall i \in \{1, \dots, K\} \\ & x_y = x_{y+1} = \dots = x_{q(y)}, \quad \forall y \in \mathcal{Y}. \end{aligned} \tag{45}$$

Proof. Consider the following optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{x}} \quad & \sum_{i=1}^{K} \frac{f_t}{x_t} \\
\text{s.t.} \quad & \sum_{t=0}^{T-1} x_t = K(b-T) \\
& x_i \geq x_{i+1}, \quad \forall i \in \{1, \dots, K-1\} \\
& x_i \geq 0, \quad \forall i \in \{1, \dots, K\}, \\
& x_y = x_{y+1} = \dots = x_{q(y)}, \quad \forall y \in \mathcal{Y}
\end{aligned} \tag{46}$$

and let $\bar{\boldsymbol{x}}^*$ be its optimal solution. Then, due to Lemma B.3, $K\tilde{\boldsymbol{x}}^* = \bar{\boldsymbol{x}}^*$. Therefore, we have that:

$$\frac{2}{K} \sum_{t=0}^{T-1} \frac{f_t}{\tilde{x}_t^* + 1} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*} = 2 \sum_{t=0}^{T-1} \frac{f_t}{\bar{x}_t^* + K} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*}$$

Furthermore, due to the fact that $\bar{x}_y^* = \bar{x}_{y+1}^* = \dots = \bar{x}_{q(y)}^*$ for all $y \in \mathcal{Y}$, we have that:

$$2 \sum_{t=0}^{T-1} \frac{f_t}{\bar{x}_t^* + K} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*} \leq 2 \sum_{t=0}^{T-1} \frac{f_t}{\bar{x}_t^*} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*}$$

At this point, we proceed by lower bounding:

$$\sum_{t=0}^{T-1} \frac{f_t}{n_t^*}.$$

More specifically, consider the following optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{x}} \quad & \sum_{i=1}^{K} \frac{f_t}{x_t} \\
\text{s.t.} \quad & \sum_{t=0}^{T-1} x_t = \Lambda \\
& x_i \geq x_{i+1}, \quad \forall i \in \{1, \dots, K-1\} \\
& x_i \geq 0, \quad \forall i \in \{1, \dots, K\}, \\
& x_y = x_{y+1} = \dots = x_{q(y)}, \quad \forall y \in \mathcal{Y},
\end{aligned} \tag{47}$$

and let $\boldsymbol{x}^*$ be its optimal solution. Then, we have that:

$$\sum_{t=0}^{T-1} \frac{f_t}{n_t^*} \geq \sum_{t=0}^{T-1} \frac{f_t}{x_t^*}. \tag{48}$$

To prove Equation (48), it is sufficient to drop the integer constraints from the (4), then, due to Lemma B.2, we can impose the equality constraints on the resulting optimization problem, and finally, we enlarge the feasible region by setting the constraints $x_i \geq 0$.

At this point, we have:

$$\frac{2}{K} \sum_{t=0}^{T-1} \frac{f_t}{\tilde{x}_t^* + 1} - \sum_{t=0}^{T-1} \frac{f_t}{n_t^*} \leq 2 \sum_{t=0}^{T-1} \frac{f_t}{\bar{x}_t^*} - \sum_{t=0}^{T-1} \frac{f_t}{x_t^*}. \tag{49}$$

By Lemma B.3, we have that:

$$\bar{x}_t^* = \frac{K(b-T)}{\Lambda} x_t^* = \frac{K(b-T)}{Kb} x_t^* = \frac{b-T}{b} x_t^* = \frac{cT-T)}{cT} x_t^* = \frac{c-1}{c} x_t^*$$

Plugging this result into Equation (49), we obtain:

$$\frac{2}{K}\sum_{t=0}^{T-1}\frac{f_t}{\tilde{x}_t^*+1}-\sum_{t=0}^{T-1}\frac{f_t}{n_t^*}\leq\left(\frac{2c}{c-1}-1\right)\sum_{t=0}^{T-1}\frac{f_t}{x_t^*}$$
$$=\frac{c+1}{c-1}\sum_{t=0}^{T-1}\frac{f_t}{x_t^*}.$$

□

At this point, we are ready to prove Theorem 4.1.

Theorem 4.1. *Let $\boldsymbol{n}^*$ be the optimal solution of problem* (4)*, f_t as in Equation* (3)*, $b\geq 2T$ and $\beta=\frac{6(T+T^2)\Lambda K}{\delta}$. Consider the DCS $\hat{\boldsymbol{n}}$ computed by Algorithm 1. Then, with probability at least $1-\delta$ it holds that:*

$$\mathbb{V}\mathrm{ar}_{\hat{\boldsymbol{n}}}\left[\hat{J}_{\hat{\boldsymbol{n}}}(\pi)\right]\leq 192\left(\frac{b}{\Lambda}\right)^{\frac{3}{2}}\log(\beta)\left(\sum_{t=0}^{T-1}\gamma^t\right)^2+4\mathbb{V}\mathrm{ar}_{\boldsymbol{n}^*}\left[\hat{J}_{\boldsymbol{n}^*}(\pi)\right]+\frac{2b}{\Lambda}\sum_{t:f_t<0}|f_t|. \quad (8)$$

Proof. From Lemma B.10, we can upper bound Equation (31) with:

$$\begin{aligned}&\frac{2}{K^2}\sum_{i=1}^{K}\sum_{t=0}^{T-1}f_t\left(\frac{1}{\bar{n}_{t,i}}-\frac{1}{\tilde{x}_t^*+1}\right)+\frac{2}{K^2}\sum_{i=1}^{K}\sum_{t:f_t<0}\frac{|f_t|}{\bar{n}_{t,i}}-\sum_{t:f_t<0}\frac{|f_t|}{\sum_{i=1}^{k}\hat{n}_{t,i}}\\&+\frac{2}{K}\sum_{t=0}^{T-1}\frac{f_t}{\tilde{x}_t^*+1}-\sum_{t=0}^{T-1}\frac{f_t}{n_t^*}.\end{aligned} \quad (50)$$

At this point, we notice that:

$$\frac{2}{K^2}\sum_{i=1}^{K}\sum_{t:f_t<0}\frac{|f_t|}{\bar{n}_{t,i}}-\sum_{t:f_t<0}\frac{|f_t|}{\sum_{i=1}^{k}\hat{n}_{t,i}}\leq\frac{2}{K}\sum_{t:f_t<0}f_t$$

Plugging this result into Equation (50), we obtain:

$$\frac{2}{K^2}\sum_{i=1}^{K}\sum_{t=0}^{T-1}f_t\left(\frac{1}{\bar{n}_{t,i}}-\frac{1}{\tilde{x}_t^*+1}\right)+\frac{2}{K}\sum_{t:f_t<0}|f_t|+\frac{2}{K}\sum_{t=0}^{T-1}\frac{f_t}{\tilde{x}_t^*+1}-\sum_{t=0}^{T-1}\frac{f_t}{n_t^*}. \quad (51)$$

Due to Lemma B.11, this can be further upper-bounded with:

$$192\left(\frac{b}{\Lambda}\right)^{\frac{3}{2}}\log\left(\frac{6(T+T^2)\Lambda K}{\delta}\right)\left(\sum_{t=0}^{T-1}\gamma^t\right)^2+\frac{2}{K}\sum_{t:f_t<0}|f_t|+\frac{2}{K}\sum_{t=0}^{T-1}\frac{f_t}{\tilde{x}_t^*+1}-\sum_{t=0}^{T-1}\frac{f_t}{n_t^*}.$$

Moreover, due to Lemma B.12, we can further bound the previous Equation with:

$$192\left(\frac{b}{\Lambda}\right)^{\frac{3}{2}}\log\left(\frac{6(T+T^2)\Lambda K}{\delta}\right)\left(\sum_{t=0}^{T-1}\gamma^t\right)^2+\frac{2}{K}\sum_{t:f_t<0}|f_t|+\frac{c+1}{c-1}\sum_{t=0}^{T-1}\frac{f_t}{x_t^*}, \quad (52)$$

where x_t^* is the solution of the following optimization problem:

$$\begin{aligned}\min_{\boldsymbol{x}}\quad&\sum_{i=1}^{K}\frac{f_t}{x_t}\\\text{s.t.}\quad&\sum_{t=0}^{T-1}x_t=\Lambda\\&x_i\geq x_{i+1},\quad\forall i\in\{1,\dots,K-1\}\\&x_i\geq 0,\quad\forall i\in\{1,\dots,K\},\\&x_y=x_{y+1}=\dots=x_{q(y)},\quad\forall y\in\mathcal{Y},\end{aligned} \quad (53)$$

Moreover, since:

$$\sum_{t=0}^{T-1} \frac{f_t}{n_t^*} \geq \sum_{t=0}^{T_1} \frac{f_t}{x_t^*},$$

Equation 52 reduces to:

$$192 \left(\frac{b}{\Lambda}\right)^{\frac{3}{2}} \log\left(\frac{6(T+T^2)\Lambda K}{\delta}\right) \left(\sum_{t=0}^{T-1} \gamma^t\right)^2 + \frac{2}{K} \sum_{t: f_t < 0} |f_t| + \frac{c+1}{c-1} \sum_{t=0}^{T-1} \frac{f_t}{n_t^*},$$

At this point, the results follows by noticing that:

$$\frac{c+1}{c-1} = \frac{b+T}{b-T} \leq 3,$$

and, by isolating $\mathbb{V}\mathrm{ar}_{\hat{\boldsymbol{n}}}\left[\hat{J}_{\hat{\boldsymbol{n}}}(\pi)\right]$ in Equation (31). □

B.3 Additional Technical Details

In this section, we provide additional techincal details that have been mentioned in the main text. More specifically, we provide (i) a formal description of the transformation between optimization problems and how we applied this technique in RIDO, (ii) difficulties in deriving closed-form solutions for the optimization problems of interest, (iii) a formal statement and proof of Equation (9), (iv) and theoretical evidence for the sub-optimality of non-adaptive methods whose variance cannot scale with the variance of the optimal DCS).

B.3.1 Additional Details on solving the empirical optimization problem

We begin with a more in-depth discussion of the transformation between optimization problems. Let $c_t \in \mathbb{R}$ for each $t \in \{0, \dots, T-1\}$, and define $\mathcal{Y} = \{i \in \{0, \dots, T-1\} : c_i < 0\}$. Let $y \in \mathcal{Y}$, and define $q(y)$ as the smallest integer in $\{y+1, \dots, T-1\}$ such that $\sum_{i=y}^{q(y)} c_i \geq 0$. Due to Lemma B.1 we know that, if $(c_0, \dots, c_{T-1}) = (f_0, \dots, f_{T-1})$, then $q(y)$ is always well-defined. At this point, consider the continuous relaxation of the original optimization problem, namely:

$$\begin{aligned}
\min_{\boldsymbol{n}} \quad & \sum_{t=0}^{T-1} \frac{1}{n_t} \left(\gamma^{2t} \mathbb{V}\mathrm{ar}(R_t) + 2 \sum_{t'=t+1}^{T-1} \gamma^{t+t'} \mathbb{C}\mathrm{ov}(R_t, R_{t'})\right) \\
\text{s.t.} \quad & \sum_{t=0}^{T-1} n_t = \Lambda \\
& n_t \geq n_{t+1}, \quad \forall t \in \{0, \dots, T-2\} \\
& n_t \geq 1, \quad \forall t \in \{0, \dots, T-1\}.
\end{aligned} \tag{54}$$

Due to Lemma B.1 and Lemma B.2, we know that the following optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{n}} \quad & \sum_{t=0}^{T-1} \frac{1}{n_t} \left(\gamma^{2t} \mathbb{V}\mathrm{ar}(R_t) + 2 \sum_{t'=t+1}^{T-1} \gamma^{t+t'} \mathbb{C}\mathrm{ov}(R_t, R_{t'})\right) \\
\text{s.t.} \quad & \sum_{t=0}^{T-1} n_t = \Lambda \\
& n_t \geq n_{t+1}, \quad \forall t \in \{0, \dots, T-2\} \\
& n_t \geq 1, \quad \forall t \in \{0, \dots, T-1\} \\
& n_y = n_{y+1} = \dots = n_{q(y)}, \quad \forall y \in \mathcal{Y},
\end{aligned} \tag{55}$$

has the same optimal solution of (55). At this point, to define the transformed problem it is sufficient to introduce additional variables y_i for any contiguous timesteps where $n_i = n_{i+1} = \dots = n_{i+k}$

holds for some integers i, k.[9] The optimization variables $n_i, n_{i+1}, \dots, n_{i+k+1}$ will be substituted with y_i. The objective function will be modified accordingly, namely:

$$\frac{f_i}{n_i} + \dots + \frac{f_{i+k}}{n_{i+k}},$$

is replaced with:

$$\frac{f_y + \dots + f_{q(y)}}{y_i}. \tag{56}$$

Consequently, any numerator in the resulting objective function of the transformed problem will be greater or equal than 0. It is easy to verify that, in this case, the resulting objective function is convex in the considered optimization domain. Finally, as a last remark, we notice that the constraint $\sum_{t=0}^{T-1} n_t = \Lambda$ needs to be modified. More specifically, if y_i substitutes l_i variables, then its contribution within the budget constraint summation will be given by $y_i l_i$.

As discussed in Section 4.1, in RIDO we adopt a procedure that is inspired by the aforementioned theoretical properties of the continuous relaxation of the optimization problem (4). Nevertheless, it has to be noticed that a modification needs to be taken into account when replacing exact quantities (i.e., f_t) with their estimation and exploration bonuses (which, in the following, we refer to as $\hat{f}_t$ for brevity). More specifically, in general, contrary to what highlighted in Lemma B.1 for the original objective function, when using $\hat{f}_t$ it might happen that $q(y)$ is not well-defined for every possible y. Indeed, due to the noise that is present in the estimation process, there might exists $\bar{t}$ such that $\hat{f}_{\bar{t}} < 0$ and $\sum_{t=\bar{t}}^{t'} \hat{f}_t < 0$ for all $t' > \bar{t}$. Whenever this condition is verified, we adopt the following heuristic to make the computation tractable. If $\bar{t} = 0$, then we just set the DCS of the current mini-batch to the uniform-in-the-horizon one. When $\bar{t} \neq 0$, instead, we group together $n_{\bar{t}}, \dots, n_{T-1}$ and we introduce a new variable y that will divide, in the objective function, $\hat{f}_{\bar{t}-1}$. As a final remark, however, we notice that these modifications do not impact on the theoretical properties of RIDO. Indeed, Lemma B.9, shows that, with probability at least $1 - \delta$, the aforementioned ill-conditions do not happen. As a consequence, we can study the high-probability behavior of RIDO assuming access to the solution of the transformed optimization problem discussed at the beginning of this section (that preserves the optimal solution of the continuous relaxation of (5)).

B.3.2 On closed-form solutions

We now continue by discussing the closed-form solutions of the optimization problems of interests. First of all, optimization problems (4) and (5) are integer and non-linear problems. Even neglecting the non-linear dependency on n, we remark that solving integer and linear problem is NP-hard. At this point, one might resort to study their continuous relaxations. In the following, we focus on the continuous relaxation of (4) (indeed, as noticed at the end of the previous section, the continuous relaxation of (5) requires additional effort). As mentioned above, whenever $f_t < 0$ holds for some $t \in \{0, \dots, T-1\}$, the continuous relaxation of f_t is non-convex. Nevertheless, from Lemma B.2, we know that we can always derive an equivalent convex problem (where the numerator in the objective function is always greater or equal than 0) that preserves the optimal solution. For

[9] More precisely, we notice that $n_i = n_{i+1} = \dots = n_{i+k}$ might involve multiple constraints in the formulation of (55). In this sense, we need to refer to the largest intervals in which these constraints are enforced, otherwise we might introduce multiple variables that refer to the same original optimization variable.

this reason, we now report the KKT conditions under the assumption that $f_t \geq 0$ holds for all $t \in \{0, \dots, T-1\}$.[10]

$$\begin{cases} -\frac{f_t}{n_t^2} + \eta - \xi_t - \mu_t \mathbf{1}\{t < T-1\} + \mu_{t-1}\mathbf{1}\{t > 0\} = 0 & \forall t \in \{0, \dots, T-1\} \\ \xi_t(1 - n_t) = 0 & \forall t \in \{0, \dots, T-1\} \\ \mu_t(n_{t+1} - n_t) = 0 & \forall t \in \{0, \dots, T-2\} \\ \eta(\sum_{t=0}^{T-1} n_t - \Lambda) = 0 & \\ \sum_{t=0}^{T-1} n_t - \Lambda = 0 & \\ \mu_t \geq 0 & \forall t \in \{0, \dots, T-2\} \\ \xi_t \geq 0 & \forall t \in \{0, \dots, T-1\} \end{cases} . \tag{57}$$

At this point, we notice that a similar problem has been solved in Poiani et al. [2023] for deriving a closed-form solutions that minimizes confidence intervals around the return estimator. In that situation, however, the constraints $n_t \geq n_{t+1}$ were not present since they were automatically satisfied by any optimal solution (and, consequently, they were removed from the optimization problem of interest). The main challenge in our setting is, indeed, the presence of $\mu_t(n_{t+1} - n_t)$, together with the terms related to μ_t in the first Equation of (57). These additional components within (57) prevented us to derive a closed-form solutions of the continuous relaxation (4) (and, (5)).

B.3.3 Proof of Equation (9)

We now continue with providing a formal statement and proof of Equation (9).

Corollary B.13. *Suppose that $f_t \geq 0$ for all $t \in \{0, \dots, T-1\}$, and $\sum_{t=0}^{T-1} f_t > 0$. Let:*

$$\Lambda \geq \Lambda_0 := \left(55296 \frac{\frac{b^{\frac{3}{2}}}{\delta} \left(\sum_{t=0}^{T-1} \gamma^t \right)^2}{\sum_{t=0}^{T-1} f_t} \right)^3 . \tag{58}$$

Let $\beta = \frac{6(T+T^2)\Lambda K}{\delta}$. Then, with probability at least $1-\delta$, it holds that:

$$\mathbb{V}\mathrm{ar}_{\hat{\boldsymbol{n}}} \left[\hat{J}_{\hat{\boldsymbol{n}}}(\pi) \right] \leq 5 \mathbb{V}\mathrm{ar}_{\boldsymbol{n}^*} \left[\hat{J}_{\boldsymbol{n}^*}(\pi) \right] . \tag{59}$$

Proof. The proof follows by analyzing, under the condition provided by Equation (58), the upper bound provided in Theorem 4.1. More specifically, since $f_t \geq 0$ holds, Theorem 4.1 reduces to:

$$\mathbb{V}\mathrm{ar}_{\hat{\boldsymbol{n}}} \left[\hat{J}_{\hat{\boldsymbol{n}}}(\pi) \right] \leq 192 \left(\frac{b}{\Lambda} \right)^{\frac{3}{2}} \log(\beta) \left(\sum_{t=0}^{T-1} \gamma^t \right)^2 + 4 \mathbb{V}\mathrm{ar}_{\boldsymbol{n}^*} \left[\hat{J}_{\boldsymbol{n}^*}(\pi) \right] .$$

To prove Equation (59) it is thus sufficient to show that, under $\Lambda \geq \Lambda_0$, the following holds:

$$192 \left(\frac{b}{\Lambda} \right)^{\frac{3}{2}} \log(\beta) \left(\sum_{t=0}^{T-1} \gamma^t \right)^2 \leq \mathbb{V}\mathrm{ar}_{\boldsymbol{n}^*} \left[\hat{J}_{\boldsymbol{n}^*}(\pi) \right] . \tag{60}$$

We proceed by lower bounding the right hand side of Equation (60).

$$\mathbb{V}\mathrm{ar}_{\boldsymbol{n}^*} \left[\hat{J}_{\boldsymbol{n}^*}(\pi) \right] = \sum_{t=0}^{T-1} \frac{f_t}{n_t^*} \geq \sum_{t=0}^{T-1} \frac{f_t}{\Lambda} = \frac{1}{\Lambda} \sum_{t=0}^{T-1} f_t,$$

where, the inequality follows from the fact that $n_t^* \leq \Lambda$ and $f_t \geq 0$. Given this result, Equation (60) holds whenever the following holds:

$$192 \left(\frac{b}{\Lambda} \right)^{\frac{3}{2}} \log(\beta) \left(\sum_{t=0}^{T-1} \gamma^t \right)^2 \leq \frac{1}{\Lambda} \sum_{t=0}^{T-1} f_t. \tag{61}$$

[10] Under the assumption that $f_t \geq 0$ holds, the problem is convex, and the KKT conditions provides necessary and sufficient conditions for optimality.

Therefore, we now focus on Equation (61), and proceed by upper-bounding its left hand side. More specifically, we have that:

$$
\begin{aligned}
192\left(\frac{b}{\Lambda}\right)^{\frac{3}{2}}\log\left(\frac{6(T+T^2)\Lambda K}{\delta}\right)\left(\sum_{t=0}^{T-1}\gamma^t\right)^2 &\leq 192\left(\frac{b}{\Lambda}\right)^{\frac{3}{2}}\log\left(\frac{12\Lambda^4}{\delta}\right)\left(\sum_{t=0}^{T-1}\gamma^t\right)^2 \\
&\leq 768\left(\frac{b}{\Lambda}\right)^{\frac{3}{2}}\log\left(\frac{12\Lambda}{\delta}\right)\left(\sum_{t=0}^{T-1}\gamma^t\right)^2 \\
&\leq 55296\frac{b^{\frac{3}{2}}}{\delta}\Lambda^{-\frac{4}{3}}\left(\sum_{t=0}^{T-1}\gamma^t\right)^2,
\end{aligned}
$$

where, in the first step we have used $T \leq \Lambda$ and $K \leq \Lambda$, in the second one we have used logarithm properties and in the last one we have used $\log x \leq \frac{x^{\frac{1}{6}}}{\frac{1}{6}}$. At this point, Equation (61) holds whenever the following holds:

$$
55296\frac{b^{\frac{3}{2}}}{\delta}\Lambda^{-\frac{4}{3}}\left(\sum_{t=0}^{T-1}\gamma^t\right)^2 \leq \frac{1}{\Lambda}\sum_{t=0}^{T-1} f_t,
$$

which can be rewritten as:

$$
\Lambda^{\frac{1}{3}} \geq 55296\frac{\frac{b^{\frac{3}{2}}}{\delta}\left(\sum_{t=0}^{T-1}\gamma^t\right)^2}{\sum_{t=0}^{T-1} f_t} \tag{62}
$$

Equation (62) is clearly satisfied for $\Lambda \geq \Lambda_0$, thus concluding the proof. □

B.3.4 Theoretical Sub-optimality of pre-determined schedules

Finally, we conclude by providing theoretical evidence on the reasons why claims similar to the one of Corollary (B.13) does not hold for pre-determined schedules (i.e, the uniform-in-the-horizon one and the robust DCS of Poiani et al. [2023]).

Proposition B.14 (Sub-optimality of the Uniform Strategy)**.** *Let $f_0 \neq 0$ and $f_i = 0$ for all $i \geq 1$. Let $T > 2$. Let $\boldsymbol{n}_u = \left(\frac{T}{\Lambda}, \dots, \frac{T}{\Lambda}\right)$. For any value of budget Λ, it does not exist a universal constant $c > 0$ for which the following holds:*

$$
\mathbb{V}\mathrm{ar}_{\boldsymbol{n_u}}\left[\hat{J}_{\boldsymbol{n_u}}(\pi)\right] \leq c\mathbb{V}\mathrm{ar}_{\boldsymbol{n^*}}\left[\hat{J}_{\boldsymbol{n^*}}(\pi)\right]. \tag{63}
$$

Proof. Under the assumption that $f_0 \neq 0$ and $f_i = 0$ for all $i \geq 1$, we have that:

$$
\mathbb{V}\mathrm{ar}_{\boldsymbol{n_u}}\left[\hat{J}_{\boldsymbol{n_u}}(\pi)\right] = \frac{T}{\Lambda}f_0, \tag{64}
$$

and, from Theorem 3.1:

$$
\mathbb{V}\mathrm{ar}_{\boldsymbol{n^*}}\left[\hat{J}_{\boldsymbol{n^*}}(\pi)\right] = \frac{1}{\Lambda-(T-1)}f_0. \tag{65}
$$

Furthermore, if $T > 2$, the variance of the optimal DCS can be upper bounded by:

$$
\mathbb{V}\mathrm{ar}_{\boldsymbol{n^*}}\left[\hat{J}_{\boldsymbol{n^*}}(\pi)\right] = \frac{1}{\Lambda-(T-1)}f_0 \leq \frac{2}{\Lambda}f_0. \tag{66}
$$

At this point, proceed by contradiction and suppose that Equation (63) holds. Then, it follows that the following equation should holds as well for some universal constant c:

$$
\frac{T}{\Lambda}f_0 \leq c\frac{2}{\Lambda}f_0. \tag{67}
$$

Equation (67) reduces to:

$$
c \geq \frac{T}{2}, \tag{68}
$$

which contradicts the claim, thus concluding the proof. □

Proposition B.15 (Sub-optimality of the Robust Strategy of Poiani et al. [2023])**.** *Let $f_0 \neq 0$ and $f_i = 0$ for all $i \geq 1$. Let $\tilde{\boldsymbol{n}}$ be the robust DCS of Poiani et al. [2023]. Let $T > 2$ and $d_t = \frac{\gamma^t(\gamma^t+\gamma^{t+1}-2\gamma^T)}{1-\gamma}$ and suppose that $\Lambda \geq \Lambda_0 := \frac{\sum_{t=0}^{T-1}\sqrt{d_t}}{\sqrt{d_{T-1}}}$. For any value of budget $\Lambda \geq \Lambda_0$, it does not exist a universal constant $c > 0$ for which the following holds*

$$\mathbb{V}\mathrm{ar}_{\tilde{\boldsymbol{n}}}\left[\hat{J}_{\tilde{\boldsymbol{n}}}(\pi)\right] \leq c\mathbb{V}\mathrm{ar}_{\boldsymbol{n}^*}\left[\hat{J}_{\boldsymbol{n}^*}(\pi)\right]. \tag{69}$$

Proof. Under the assumption that $f_0 \neq 0$, $f_i = 0$ for all $i \geq 1$, and $\Lambda \geq \Lambda_0$ we have that:[11]

$$\mathbb{V}\mathrm{ar}_{\tilde{\boldsymbol{n}}}\left[\hat{J}_{\tilde{\boldsymbol{n}}}(\pi)\right] \geq \frac{f_0}{2\Lambda}\frac{\sum_{t=0}^{T-1}\sqrt{d_t}}{\sqrt{d_0}}. \tag{70}$$

From Theorem 3.1:

$$\mathbb{V}\mathrm{ar}_{\boldsymbol{n}^*}\left[\hat{J}_{\boldsymbol{n}^*}(\pi)\right] = \frac{1}{\Lambda-(T-1)}f_0. \tag{71}$$

Furthermore, if $T > 2$, the variance of the optimal DCS can be upper bounded by:

$$\mathbb{V}\mathrm{ar}_{\boldsymbol{n}^*}\left[\hat{J}_{\boldsymbol{n}^*}(\pi)\right] = \frac{1}{\Lambda-(T-1)}f_0 \leq \frac{2}{\Lambda}f_0. \tag{72}$$

At this point, proceed by contradiction and suppose that Equation (69) holds. Then, it follows that the following equation should hold as well for some universal constant c:

$$\frac{f_0}{\Lambda}\frac{\sum_{t=0}^{T-1}\sqrt{d_t}}{\sqrt{d_0}} \leq c\frac{2}{\Lambda}f_0. \tag{73}$$

Equation (73) can be rewritten as:

$$c \geq \frac{1}{2}\frac{\sum_{t=0}^{T-1}\sqrt{d_t}}{\sqrt{d_0}}. \tag{74}$$

However, if Equation (74) holds, then, also the following holds:

$$c \geq \frac{1}{4}\sum_{t=0}^{T-1}\sqrt{\gamma^t\left(\gamma^t+\gamma^{t+1}-\gamma^{2T}\right)}, \tag{75}$$

which, however, contradicts the claim [12], thus concluding the proof. □

Proposition B.14 and B.15 shows that the variance of both schedules cannot attain the minimum variance up to multiplicative constant factors as RIDO, instead, does (notice, indeed, that the assumptions on f_t fits the ones of Corollary B.13). These results complements, in this sense, what has been presented in the main text, and highlights the theoretical benefits of adaptive DCSs.

C Experiment Details and Additional Results

In this section, we provide further details on our experimental settings and additional results. Section C.1 contains descriptions on the environments, Section C.2 contains details regarding hyperparameters, and Section C.3 contains additional results.

Our results have been produced using 100 Intel(R) Xeon(R) Gold 6238R CPU @ 2.20GHz cpus and 256GB of RAM. The total time taken to have all the results is around 2 weeks of computation.

C.1 Environment Details

In this section, we provide additional details on the environments that we used in our experiments.

[11] Notice that the requirement $\Lambda \geq \Lambda_0$ provides a simple closed-form expression for the robust DCS of Poiani et al. [2023]. The reader can refer to Theorem 3.3 and Appendix B of Poiani et al. [2023].

[12] Indeed, it is sufficient to take $T \to +\infty$, and $\gamma \to 1$, to show that Equation (75) tends to $+\infty$.

Ablation Domains In Setion C.3, the reader can find results and ablations that involve the scenarios described as examples in Section 1, namely Examples 1 and 2. We now provide a precise description of these domains. We start with Example 1, where the reward is gathered only at the end of the estimation horizon T. The state space is described by a 1-dimensional vector that contains only the interaction timestep t; the action space is a discrete set $\{0, 1\}$. The agent receives reward 0 in the first 9 timesteps. In the last step, instead, it receives $r \sim \mathcal{N}(3, 10)$ for action 0, and $r \sim \mathcal{N}(2, 10)$ for action 1. Concerning Example 2, instead, the setup is identical to the one of Example 1, with the only different that the non-zero reward is receives in the first interaction step. The policy that we evaluate is the uniform random.

Continuous Navigation Here, we describe in more details the 2D continuous navigation environment that we used in our experiments. The state space $\mathcal{S}$ is 2-dimensional vector $\boldsymbol{s} = (s_0, s_1) \in \mathbb{R}^2$ such that $s_i \in [0, 92]$ for all i. Similarly, the action space $\mathcal{A}$ is a 2-dimensional vector $\boldsymbol{a} = (a_0, a_1)$ such that $a_i \in [-1, 1]$ for all i. When the agent takes action $\boldsymbol{a}$ in state $\boldsymbol{s}$, it transitions to a new state $\boldsymbol{s'}$ such that:

$$s'_0 = \max\{0, \min\{s_0 + q_0, 92\}\}, \qquad s'_1 = \max\{0, \min\{s_1 + q_1, 92\}\},$$

where $q_0 \sim \mathcal{N}(a_0, 0.1)$, $q_1 \sim \mathcal{N}(a_1, 0.1)$, and the max-min operations simply guarantees that the resulting state lies within the desired state space $\mathcal{S}$. The agent receives rewards egual to 0 at every time step, except when the resulting state $\boldsymbol{s'}$ falls within a goal region. More specifically, the goal is defined as a 2-dimensional vector $\boldsymbol{g} = (91, 91)$. Whenever $||\boldsymbol{s'} - \boldsymbol{g}||_2 \leq 1$ the reward received by the agent is sampled from the following Gaussian distribution: $\mathcal{N}(1, 1)$. The agent starts in a random position that is sampled from a uniform distribution in the area $[0, 5] \times [0, 5]$. The agent policy that we evaluate in our experiments is an hand-coded expert policy that minimizes the distance between the agent's position and the center of the goal area. More specifically, given the agent position $\boldsymbol{s}$, $\boldsymbol{a}$ is computed in the following way.

$$a_0 = \max\{-1, \min\{g_0 - s_0, 1\}\}, \qquad a_1 = \max\{-1, \min\{g_1 - s_1, 1\}\},$$

where the max-min operation guarantees that $\boldsymbol{a}$ belongs to $\mathcal{A}$.

LQG Concerning the LQG, we consider the following 1-dimensional case (i.e., the dimension of the state and action spaces is 1). The initial state is drawn from a uniform distribution in $[-80, +80]$. Upon taking action $a \in \mathcal{A}$, the agent transitions to a new state $s' = s + (a + \xi) + \eta$, where $\eta \sim \mathcal{N}(0, 0.1)$ models the noise in the system, and $\xi \sim \mathcal{N}(0, 0.1)$ denotes the controller's noise. The reward for taking action a in state s is computed as $s^2 + (a + \xi)^2$. The policy that we evaluate is the optimal one and it is computed by solving the Riccati equations.

MuJoCo suite In the main text, we presented results on the Ant environment of the MuJoCo suite. In the appendix, we present additional experiments on the HalfCheetah and Swimmer domains [Todorov et al., 2012]. In all cases, we adopted trained deep RL agents made publicly available by Raffin [2020] (MIT License).

C.2 Hyper-parameters

Table 1 reports the hyper-parameters that we used in our experiments. To select the robustness level β we tried different values in $[1, 3]$, while for the batch-size we tried different values in $[2T, 10T]$. We then report the results using the best hyper-parameters configuration.

C.3 Additional Results

C.3.1 Ablations

In this section, we present ablations on RIDO on the two environments (described in Section C.1) that models Examples 1 and 2. More specifically, we conduct the following two ablations to understand the behavior of RIDO according to changes in its hyper-parameters, i.e., the robustness level β and the mini-batch size. To properly assess the effect of these designer's choices, we report and discuss both the average variance and the resulting DCSs. We test our method using $\gamma = 1$, but similar results can be obtained varying the discount factor.

Table 1: Hyper-Parameters

Environment	β	Mini-batch size
Pendulum	1.01	500
LQG	2.0	400
2D Continuous Navigation	1.0	1000
Ant	1.5	3000
HalfCheetah	2	3000
Swimmer	1.0	1000

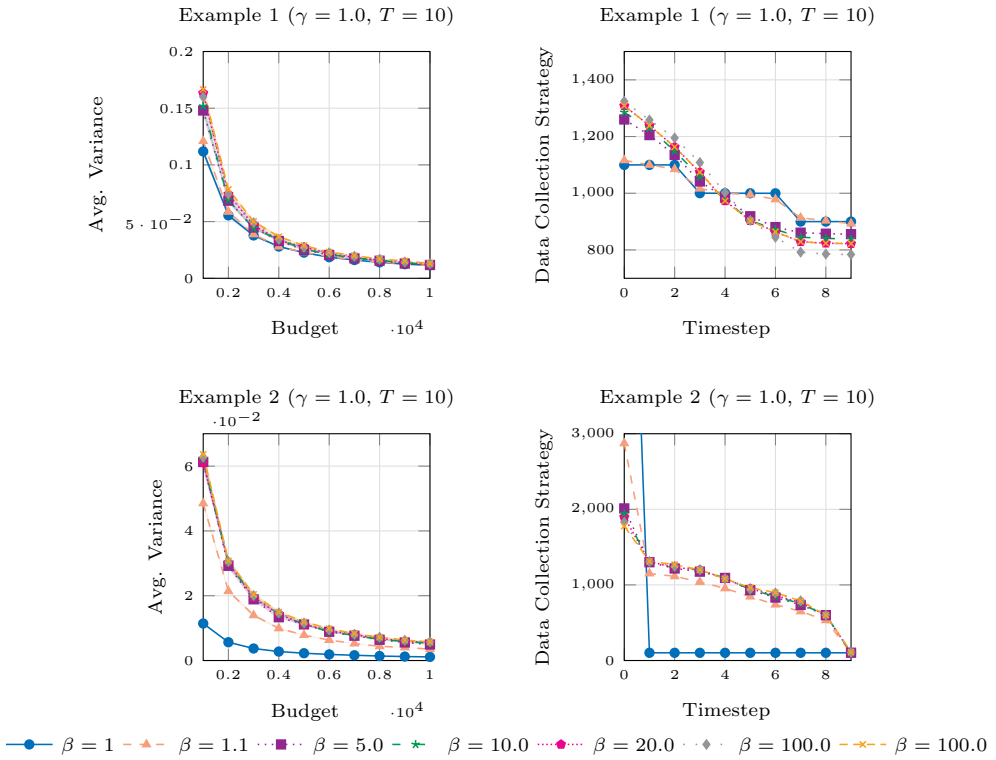

Figure 3: Ablations on different values of β on Examples 1 (*top*) and 2 (*bottom*). Empirical variance (mean and 95% confidence intervals over 100 runs) (*left*). DCS visualiaztion (mean and 95% confidence intervals over 100 runs) using $\Lambda = 10000$ (*right*).

Ablations on β We begin by performing an ablation on the robustness parameter β. More specifically, we analyze the behavior of RIDO for the following values of β: $\{1, 1.1, 5.0, 10.0, 20.0, 100.0, 1000.0\}$ (the value of the mini-batch size here is fixed to $b = 100$). Figure 3 reports the results. Let us first focus on Example 1 (i.e., the top row). In this case, the reward is gathered at the end of the estimation horizon. As we can see, increasing the value of β, leads to a larger amount of data spent in the first interaction steps (i.e., top-right in Figure 3). Indeed, when higher values of β are used, the cumulative sum of exploration bonuses in the early steps is larger w.r.t. the late ones. For this reason, RIDO spends a larger portion of its budget to decrease these exploration bonuses. As a consequence, given that the reward process of the underlying environment, this results in a higher empirical variance (i.e., top-left in Figure 3). Furthermore, given that the reward is 0 everywhere except at $t = T - 1$, even using the smallest value of β (i.e., $\beta = 1$) allows the algorithm to quickly adapt its DCS toward the most relevant timestep (i.e., $t = T - 1$). Similar comments hold for Example 2 as well (i.e., bottom row in Figure 3). Finally, we notice that the behavior changes almost unsgnificantly for values of β larger than 5.0.

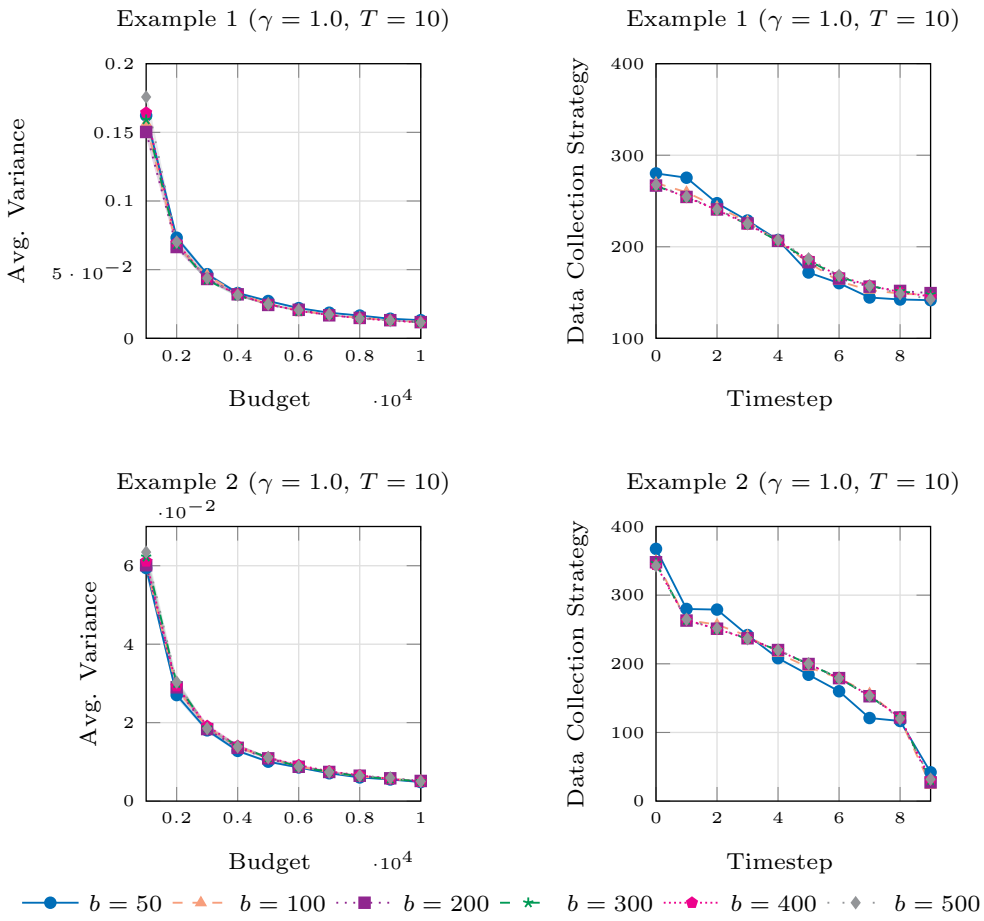

Figure 4: Ablations on different mini-batch sizes on Examples 1 (*top*) and 2 (*bottom*). Empirical variance (mean and 95% confidence intervals over 100 runs) (*left*). DCS visualiaztion (mean and 95% confidence intervals over 100 runs) using $\Lambda = 1000$ (*right*).

Ablations on mini-batch size b We now continue by presenting an ablation on the batch size. More specifically, we analyze the behavior of RIDO for the following values of b: $\{50, 100, 200, 300, 400, 500\}$ (the value of β here is fixed to 5.0). Figure 4 reports the results. First of all, as we can notice, in both Examples 1 and 2 the mini-batch size impacts the performance in a less significant way w.r.t. to the value of β (compare the left column of Figure 4 and 3). Secondly, let us focus on the the top-row (i.e., Example 1, where the reward is gathered at the end of the episode). For the smallest value of Λ of Figure 4 (i.e., $\Lambda = 1000$, that is the only for which there is some difference in performance), we notice that the best configuration is not $b = 50$ (i.e., the smallest batch-size among the presented ones). This is confirmed also by its corresponding DCS, which is not the one that allocates the highest number of data at $T - 1$. We conjecture that the reason behind this phenonema are numerical instabilities that might arise while solving the empirical problem with the use of convex solvers.[13] Concerning Example 2 (where the reward is gathered only at $t = 0$), we notice that smaller values of b performs better (this is confirmed by the corresponding DCS, that allocates more data to the first interaction step). In this case, the aformentioned problem is not present. We conjecture that the reason is that, even in the case of numerical instabilities, errors that arise from converting the continuous DCS to its integer version provably minimizes the variance, since the remaining budget is allocated uniformly starting from $t = 0$ (i.e., the most relevant timestep from

[13] We notice that even small imprecisions can result in DCSs that differ by 1 when converting the continuous relaxation to its integer version. For smaller values of b, this behavior might happen multiple times w.r.t. larger values of b.

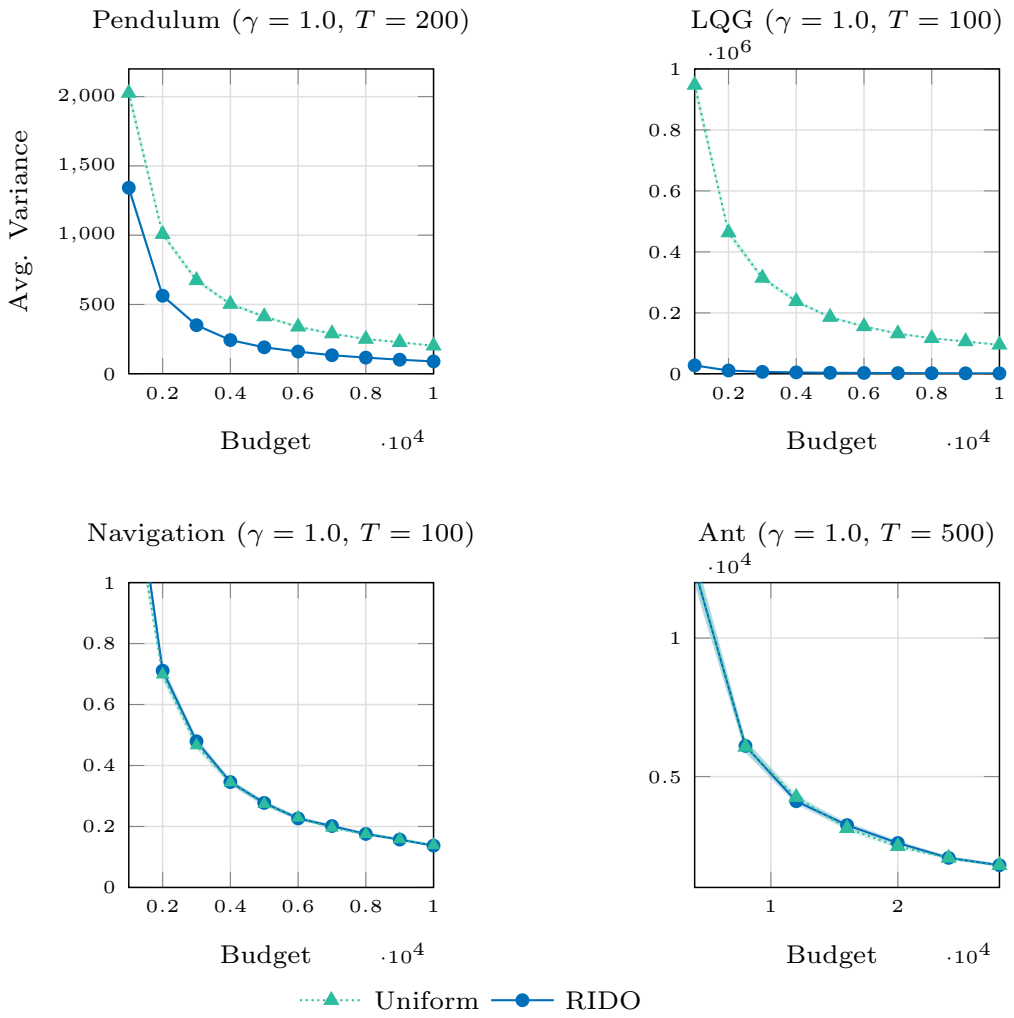

Figure 5: Empirical variance (mean and 95% confidence intervals over 100 runs) on the considered domains and baselines using $\gamma = 1$.

the point of view of the estimation quality). Finally, we notice that, whatever value of b we use, the behavior of RIDO is stable under reasonable variations of the mini-batch size.

C.3.2 Experiments with $\gamma = 1$

In this section, we present results under the experimental setting of Figure 2 but using $\gamma = 1$. Figure 5 reports the comparison between RIDO and the uniform-in-the-horizon strategy. Notice that these experiments highlight a particular beneficial feature of RIDO w.r.t. the schedule of Poiani et al. [2023]. Indeed, when $\gamma = 1$ their robust DCS does not formally exists (i.e., the method requires $\gamma < 1$, and when $\gamma \to 1$, their strategy tends to the uniform one). RIDO, on the other hand, does not heavily rely on the property of discounted sum and can be applied as-is also when $\gamma = 1$. Furthermore, Figure 5 confirms the adaptivity of RIDO that has already been highlighted in the main text. Namely, it does not underperform the uniform-in-the-horizon strategy when long trajectories are required, while it reduces the return estimator's variance when truncated trajectories are convenient.

C.3.3 DCS Visualizations for Figure 2

In this section, we present visualizations of the DCSs for the experiments presented in Figure 2 and 5. Figure 6 and 7 reports our results (mean and 95% confidence intervals over 100 runs). For $\gamma = 1$, the robust DCS is missing since it coincides with the uniform-in-the-horizon one (further details on

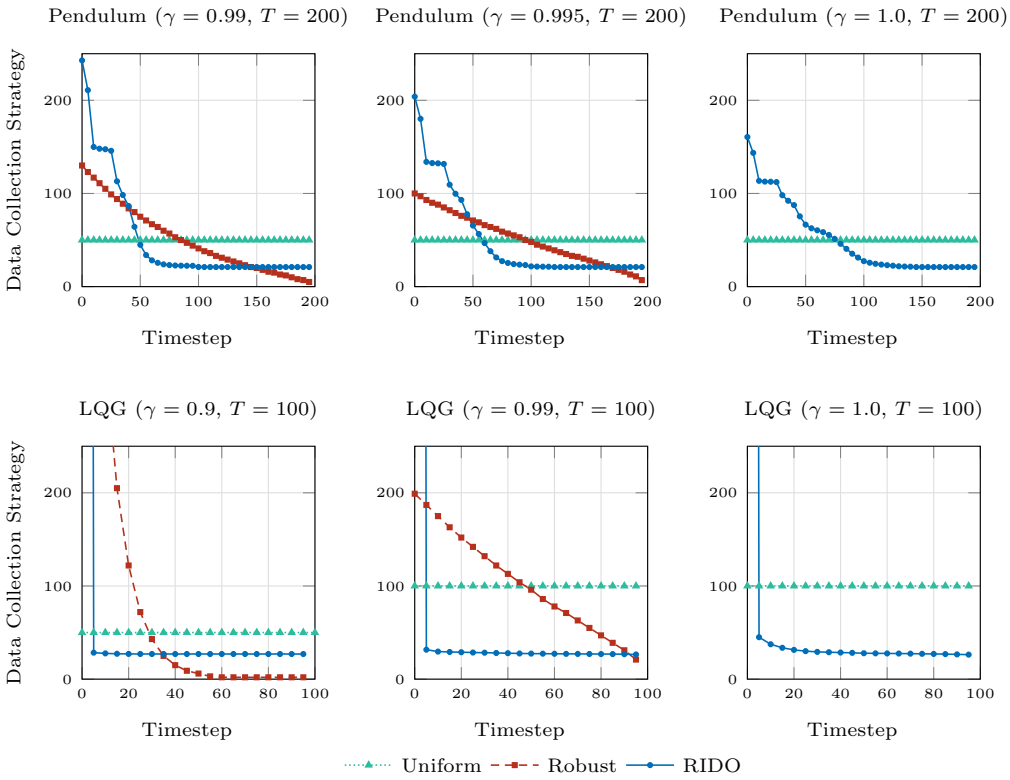

Figure 6: DCS visualiaztion for Pendulum and LQG (mean and 95% confidence intervals over 100 runs). The x axis reports the timestep t, while the y axis n_t. We consider $\Lambda = 10000$ both for the Pendulum and the LQG.

this point are available in Appendix C.3.2). The resulting visualizations reinforce the adaptivity of RIDO. Indeed, depending on the domain, the behavior of RIDO changes significantly, resulting in behaviors that are similar to the uniform strategy (i.e., Navigation), to the robust strategy (i.e., Ant), or significantly different from both pre-computed schedules (i.e., Pendulum and LQG).

C.3.4 Results on Additional Environments

In this section, we present results on additional environments from the MuJoCo suite, namely HalfCheetah and Swimmer. Figure 8 reports our results, and Figure 9 the DCSs visualization in the considered domain. In these cases, RIDO confirms its adaptivity achieving a satisfying performance level. For HalfCheetah similar comments w.r.t. made for the Ant in Figure 2. In the Swimmer domain, on the other hand, the robust DCS shows sub-optimal performance w.r.t. the uniform schedule and RIDO.

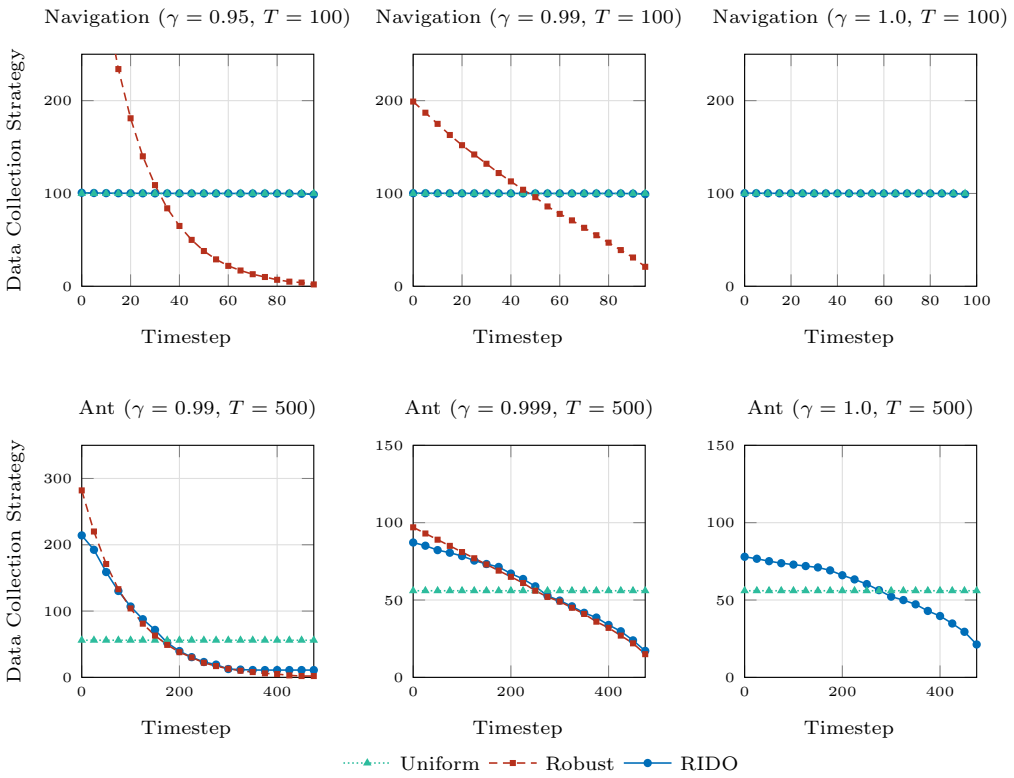

Figure 7: DCS visualization for 2D Continuous Navigation and Ant (mean and 95% confidence intervals over 100 runs). The x axis reports the timestep t, while the y axis n_t. For 2D Continuous Navigation we consider $\Lambda = 10000$, while for the Ant we consider experiments using $\Lambda = 28000$.

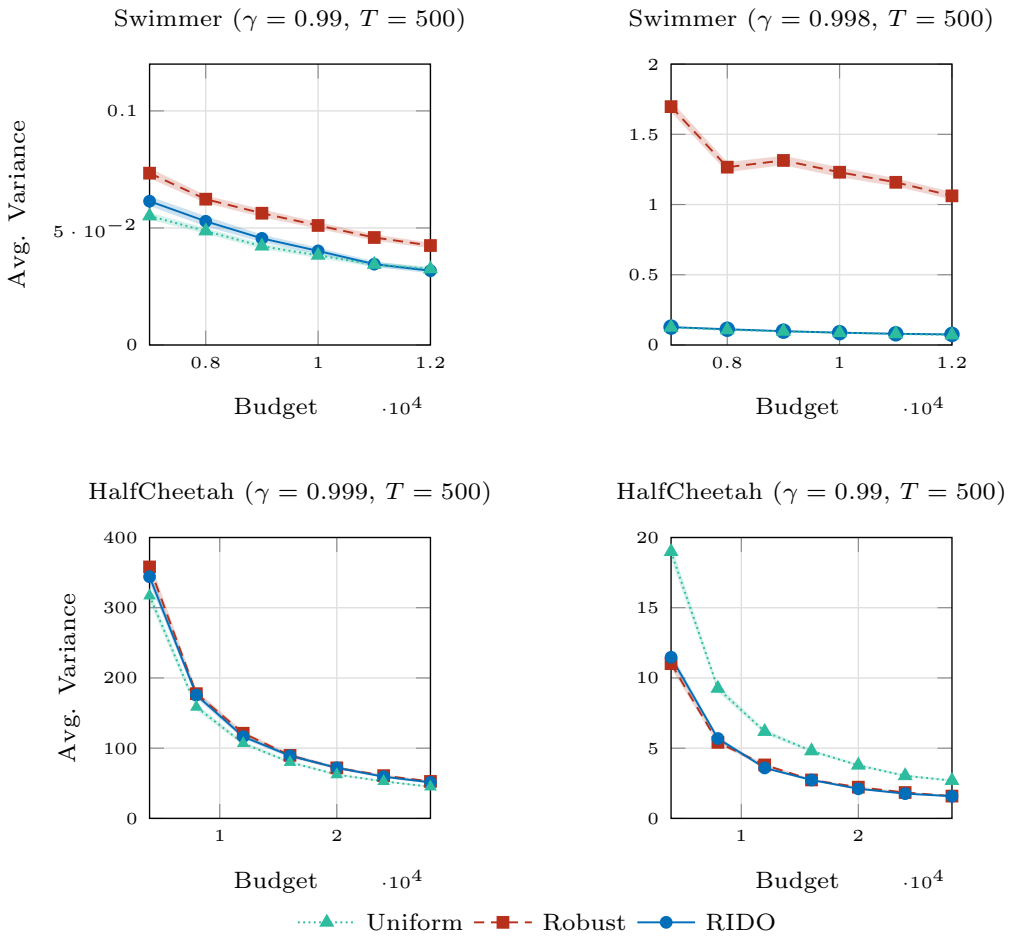

Figure 8: Empirical variance (mean and 95% confidence intervals over 100 runs) on the Swimmer and HalfCheetah for the considered baselines.

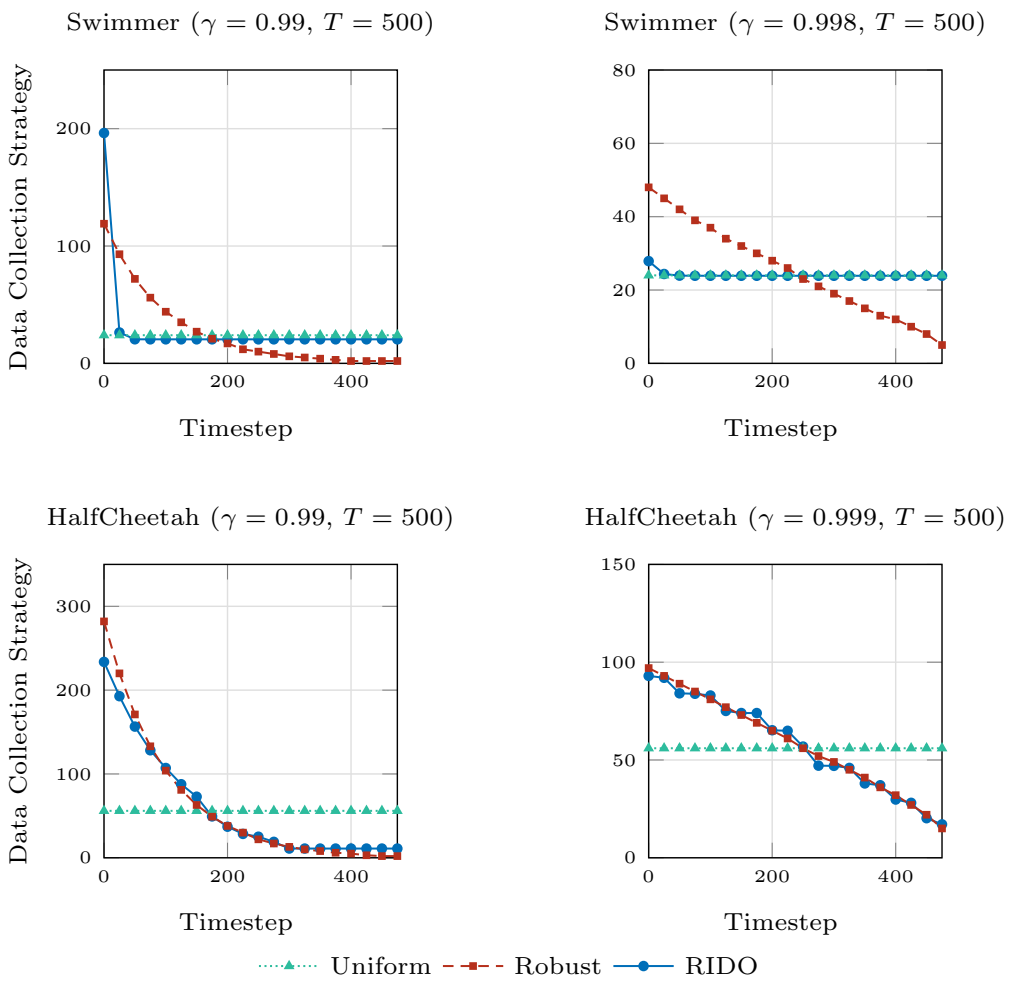

Figure 9: DCS visualization for Swimmer and HalfCheetah (mean and 95% confidence intervals over 100 runs). The x axis reports the timestep t, while the y axis n_t. For Swimmer we consider $\Lambda = 12000$, while for the HalfCheetah we consider experiments using $\Lambda = 28000$.