Reward-Free Policy Space Compression for Reinforcement Learning

Mirco Mutti [1 2] **Stefano Del Col** [1] **Marcello Restelli** [1]

Abstract

In reinforcement learning, we encode the potential behaviors of an agent interacting with an environment into an infinite set of policies, called *policy space*, typically represented by a family of parametric functions. Dealing with such a policy space is a hefty challenge, which often causes sample and computational inefficiencies. However, we argue that a limited number of policies is actually relevant when we also account for the structure of the environment and of the policy parameterization, as many of them would induce very similar interactions, i.e., state-action distributions. In this paper, we seek for a reward-free *compression* of the policy space into a finite set of representative policies, such that, given any policy π, the minimum Rényi divergence between the state-action distributions of the representative policies and the state-action distribution of π is bounded. We show that this compression of the policy space can be formulated as a set cover problem, and it is inherently NP-hard. Nonetheless, we propose a game-theoretic reformulation for which a locally optimal solution can be efficiently found by iteratively stretching the compressed space to cover the most challenging policy. Finally, we provide an empirical evaluation to illustrate the compression procedure in simple domains, and its ripple effects in reinforcement learning.

1. Introduction

In the Reinforcement Learning (RL) (Sutton & Barto, 2018) framework, an artificial agent interacts with an environment, typically modeled through a Markov Decision Process (MDP) (Puterman, 2014), to maximize some form of long-term performance, which is usually the sum of the discounted rewards collected in the process. The agent's behavior is encoded in a Markovian *policy*, i.e., a function that maps the current state of the environment with a probability distribution over the next action to be taken. In principle, if the underlying MDP is small enough, we can represent a Markovian policy with a table that includes an entry for each state-action pair, and we call it a *tabular* policy. However, most relevant scenarios have too many (possibly infinite) states and actions to allow for a tabular representation. In this case, we can turn to function approximation (Sutton & Barto, 2018) to encode the policy within a family of parametric functions, e.g., a linear basis combination or a deep neural network, and we call it a *parametric* policy. This set of parametric policies, which we call the *policy space*, is typically infinite. Therefore, learning a policy that maximizes the performance can be a hefty challenge, and the sheer size of the policy space often causes sample and computation inefficiencies.

A setting where these inefficiencies arise clearly and naturally is Policy Optimization (PO) (Deisenroth et al., 2013). In PO, we aim to find a policy that maximizes the performance within the policy space, i.e., an *optimal* policy, with the least amount of interactions (Sutton et al., 1999; Silver et al., 2014; Schulman et al., 2015; Metelli et al., 2018). If we also account for the performance of the policies that are actually deployed to collect these interactions, we come up with an online PO (Papini et al., 2019). In this setting, we try to minimize the *regret* that the agent suffers by taking interactions with a sub-optimal behavior before converging to an optimal policy. Recent results showed that the regret of online PO is directly related to the size of the policy space (Papini et al., 2019; Metelli et al., 2020a). In particular, online PO with a finite policy space can enjoy a constant regret, i.e., it does not scale with the number of interactions, under certain conditions (Metelli et al., 2020a). Instead, the regret of online PO with an infinite policy space does scale with the square root of the number of interactions in general (Papini et al., 2019), which means that we only have asymptotic guarantees of reaching an optimal policy. In view of these results, one could wonder whether the expressive power of an infinite policy space is worth the additional regret it causes: Are all of these infinitely many policies really necessary for PO? The expressive power of a policy space is related to the different distributions that its policies can induce over the states and actions of the environment, as the whole point of PO is to find a policy

[1] Politecnico di Milano, Milan, Italy [2] Università di Bologna, Bologna, Italy. Correspondence to: Mirco Mutti <mirco.mutti@polimi.it>.

Unsupervised Reinforcement Learning Workshop at ICML 2021.

that maximizes the probability of reaching state-action pairs associated with high rewards. However, different parameterizations might actually induce equivalent policies due to the specific structure of the policy space. Similarly, even different policies can induce the same state-action distribution in a given environment. These two types of policies are arguably redundant for PO and we would like to find a policy space that does not include either. Especially, we aim to answer the following question:

Having an infinite parametric policy space Θ in a given environment $\mathcal{M}$, can we compress Θ into a finite subset that retains most of its expressive power?

In this paper, we formulate this question into the *Policy Space Compression* problem, where we exploit the inherent structure of $\mathcal{M}$ and Θ to compute the compressed policy space. The general idea is to identify a finite set of representative policies, such that for any policy π of the original space, the minimum Rényi divergence between the state-action distributions of the representative policies and the state-action distribution of π is bounded by a given constant. This compression is agnostic to the reward function, and thus the resulting policy space can benefit the computational and sample complexity of any RL task one can later specify over $\mathcal{M}$, as it is typical in reward-free RL (Hazan et al., 2019; Jin et al., 2020a).

Especially, the paper includes the following contributions. First, we provide a formal definition of the policy space compression problem (Section 3). We note that the problem can be formulated equivalently as a set cover, and that finding an optimal compression of the policy space is NP-hard in general (Feige, 1998). Despite this negative result, we propose a game-theoretic reformulation (Section 4) that casts the problem to the one of reaching a differential Stackelberg equilibrium (Fiez et al., 2020) of a two-player sequential game, in which the first player tries to cover the policy space with a finite set of policies and the second player tries to find a policy that falls outside this coverage. Then, we present an algorithm (Section 5) to efficiently compute a compression of the policy space by repeatedly solving, with a first-order method, the two-player game for an increasing number of covering policies, until the compression requirement is met globally. In Section 6, we provide a theoretical analysis of the performance guarantees attained by the compressed policy space in relevant RL tasks. Finally, in Section 7 we provide a brief numerical validation of both the compression algorithm and RL with the compressed policy space. The proofs of the theorems can be found in Appendix A.

2. Preliminaries

In this section, we introduce the essential background on controlled Markov processes, policy optimization, importance sampling estimation and its relationship with the Rényi divergence.

2.1. Controlled Markov Processes

A discrete-time Controlled Markov Process (CMP) is defined as a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, \mu, \gamma)$, in which $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition model such that the next state is drawn as $s' \sim P(\cdot|s,a)$ given the current state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, $\mu : \Delta(\mathcal{S})$ is an initial state distribution such that the initial state is drawn as $s \sim \mu(\cdot)$, and $\gamma \in (0,1]$ is the discount factor. The behavior of an agent interacting with a CMP can be modeled through a Markovian parametric policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta(\mathcal{A})$ such that an action is drawn as $a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)$ given the current state $s \in \mathcal{S}$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^m$ are the policy parameters, and the set Π_Θ is called the *policy space*. A policy $\pi_{\boldsymbol{\theta}}$ induces a γ-discounted state distribution $d^s_{\pi_{\boldsymbol{\theta}}} : \Delta(\mathcal{S})$ over the state space of the CMP $\mathcal{M}$, which is given by $d^s_{\pi_{\boldsymbol{\theta}}}(s) = (1-\gamma)\sum_{t=1}^{\infty} \gamma^t Pr(s_t = s)$ or the equivalent recursive relation $d^s_{\pi_{\boldsymbol{\theta}}}(s) = (1-\gamma)\mu(s) - \gamma \int_{\mathcal{SA}} d^s_{\pi_{\boldsymbol{\theta}}}(s')\pi_{\boldsymbol{\theta}}(a'|s')P(s|s',a') \,\mathrm{d}s' \,\mathrm{d}a'$. Similarly, we define the γ-discounted state-action distribution $d^{sa}_{\pi_{\boldsymbol{\theta}}} : \Delta(\mathcal{S} \times \mathcal{A})$ given by $d^{sa}_{\pi_{\boldsymbol{\theta}}}(s,a) = \pi_{\boldsymbol{\theta}}(a|s) d^s_{\pi_{\boldsymbol{\theta}}}(s)$. With a slight overloading of notation, we will indifferently denote the parametric policy space Π_Θ by Θ, a parametric policy $\pi_{\boldsymbol{\theta}} \in \Pi_\Theta$ by $\boldsymbol{\theta}$, and its induced distributions $d^s_{\pi_{\boldsymbol{\theta}}}(s), d^{sa}_{\pi_{\boldsymbol{\theta}}}(s,a)$ by $d^s_{\boldsymbol{\theta}}(s), d^{sa}_{\boldsymbol{\theta}}(s,a)$.

2.2. Policy Optimization

The process of looking for the policy that maximizes the agent's performance on a given RL task with a direct search in the policy space is called Policy Optimization (PO) (Deisenroth et al., 2013). The task is generally modeled through a Markov Decision Process (MDP) (Puterman, 2014) $\mathcal{M}^{\mathcal{R}} := \mathcal{M} \cup \mathcal{R}$, i.e., the combination of a CMP $\mathcal{M}$ and a reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [-\mathrm{R}_{\max}, \mathrm{R}_{\max}]$ such that $R(s,a)$ is the bounded reward that the agent collects by selecting action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, and $\mathrm{R}_{\max} < \infty$. The agent's performance is defined by the expected sum of discounted rewards collected by its policy, i.e.,

$$\begin{aligned} J(\boldsymbol{\theta}) &:= \mathop{\mathbb{E}}_{\substack{s_0 \sim \mu(\cdot) \\ a_t \sim \pi_{\boldsymbol{\theta}}(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t,a_t)}} \left[\sum_{t=1}^{\infty} \gamma^t \mathcal{R}(s_t,a_t) \right] \\ &= \frac{1}{(1-\gamma)} \mathop{\mathbb{E}}_{(s,a) \sim d^{sa}_{\boldsymbol{\theta}}} \big[\mathcal{R}(s,a)\big], \end{aligned}$$

A Monte-Carlo estimate of the performance can be computed from a batch of N samples $\{s_n, a_n\}_{n=1}^N$ taken with the policy $\pi_{\boldsymbol{\theta}}$ in the γ-discounted MDP $\mathcal{M}^{\mathcal{R}}$ as $\widehat{J}(\boldsymbol{\theta}) = \frac{1}{(1-\gamma)N} \sum_{n=1}^N \mathcal{R}(s_n,a_n)$.

2.3. Importance Sampling and Rényi Divergence

Importance Sampling (IS) (Cochran, 2007; Owen, 2013) is a common technique to estimate the expectation of a function under a *target* distribution by taking samples from a different distribution. In PO, importance sampling allows for estimating the performance of a target policy $\pi_{\boldsymbol{\theta}'}$ through a batch of samples $\{s_n, a_n\}_{n=1}^N$ taken with a policy $\pi_{\boldsymbol{\theta}}$. Especially, we define the importance weight $w_{\boldsymbol{\theta}'/\boldsymbol{\theta}}(s,a) := d^{sa}_{\boldsymbol{\theta}'}(s,a)/d^{sa}_{\boldsymbol{\theta}}(s,a)$. A Monte-Carlo estimate of $J(\boldsymbol{\theta}')$ via importance sampling is given by

$$\widehat{J}_{IS}(\boldsymbol{\theta}'/\boldsymbol{\theta}) = \frac{1}{(1-\gamma)N}\sum_{n=1}^{N} w_{\boldsymbol{\theta}'/\boldsymbol{\theta}}(s_n,a_n)\mathcal{R}(s_n,a_n).$$

The latter estimator $\widehat{J}_{IS}(\boldsymbol{\theta}'/\boldsymbol{\theta})$ is known to be unbiased, i.e., $\mathbb{E}_{\boldsymbol{\theta}}[\widehat{J}_{IS}(\boldsymbol{\theta}'/\boldsymbol{\theta})] = J(\boldsymbol{\theta}')$ (Owen, 2013). However, $\widehat{J}_{IS}(\boldsymbol{\theta}'/\boldsymbol{\theta})$ might suffer from a large variance whenever the importance weights $w_{\boldsymbol{\theta}'/\boldsymbol{\theta}}(s,a)$ have a large variance. The variance of the importance weights is related to the exponentiated 2-Rényi divergence $D_2(d^{sa}_{\boldsymbol{\theta}'}||d^{sa}_{\boldsymbol{\theta}})$ (Rényi et al., 1961) through $\mathbb{V}\mathrm{ar}_{(s,a)\sim d^{sa}_{\boldsymbol{\theta}}}[w_{\boldsymbol{\theta}'/\boldsymbol{\theta}}(s,a)] = D_2(d^{sa}_{\boldsymbol{\theta}'}||d^{sa}_{\boldsymbol{\theta}}) - 1$ (Cortes et al., 2010), where

$$D_2(d^{sa}_{\boldsymbol{\theta}'}||d^{sa}_{\boldsymbol{\theta}}) := \int_{\mathcal{SA}} d^{sa}_{\boldsymbol{\theta}}(s,a)\left(\frac{d^{sa}_{\boldsymbol{\theta}'}(s,a)}{d^{sa}_{\boldsymbol{\theta}}(s,a)}\right)^2 \mathrm{d}s\,\mathrm{d}a.$$

The last result has been employed in (Metelli et al., 2018) to upper bound the variance of the importance sampling estimator $\widehat{J}_{IS}(\boldsymbol{\theta}'/\boldsymbol{\theta})$ as $\mathbb{V}\mathrm{ar}_{(s,a)\sim d^{sa}_{\boldsymbol{\theta}}}[\widehat{J}_{IS}(\boldsymbol{\theta}'/\boldsymbol{\theta})] \leq \left(\frac{\mathrm{R_{max}}}{1-\gamma}\right)^2 D_2(d^{sa}_{\boldsymbol{\theta}'}||d^{sa}_{\boldsymbol{\theta}})/N$. In the following, we will refer to the exponentiated 2-Rényi divergence as the Rény divergence.

3. The Policy Space Compression Problem

Let us suppose to have a CMP $\mathcal{M}$ the agent can interact with, and a parametric policy space Θ from which the agent can select its strategy of interaction. For the common parameterization choices, ranging from linear policies to deep neural networks, the policy space Θ is typically infinite. Dealing with such a large policy space to address the usual RL tasks, e.g., finding a convenient task-agnostic sampling strategy (Hazan et al., 2019) or seeking for an optimal policy within the set (Deisenroth et al., 2013), it is often a huge challenge. Furthermore, many policies in Θ are unnecessary for these purposes, as they induce very similar interactions, and thus they have very similar performance. On the one hand, different policy parameters $\boldsymbol{\theta} \in \Theta$ might induce nearly identical distributions over actions. On the other hand, even different distributions over actions can lead to comparable state-action distributions due to the structure of the environment. Since we do not have any reward encoded in $\mathcal{M}$, it would be unwise to deem any state-action distribution irrelevant without additional information on the task structure. In this work, we aim to identify a subset of the policy space $\Theta' \subseteq \Theta$ that retains most of the expressive power of Θ, i.e., the set of the state-action distributions it can induce, while dramatically reducing its size, to the advantage of the computational and sample efficiency of future RL tasks. Especially, we consider a σ-soft compression of Θ, where for any policy $\boldsymbol{\theta} \in \Theta$ we would like to have a policy $\boldsymbol{\theta}' \in \Theta'$ such that the Rényi divergence between their respective state-action distributions $d^{sa}_{\boldsymbol{\theta}}, d^{sa}_{\boldsymbol{\theta}'}$ is bounded by a positive constant σ. The Rényi divergence is particularly convenient in this setting due to its relationship with the variance of the importance sampling in the off-policy estimation (Cortes et al., 2010; Metelli et al., 2018). The following statement provides a more formal definition of this σ-soft compression.

Definition 3.1 (σ-compression)**.** *Let $\mathcal{M}$ be a CMP, let Θ be a parametric policy space for $\mathcal{M}$, and let $\sigma > 0$ be a constant. We call Θ_σ a σ-compression of Θ in $\mathcal{M}$ if it holds that $|\Theta_\sigma| < \infty$ and*

$$\forall \boldsymbol{\theta} \in \Theta, \quad \min_{\boldsymbol{\theta}' \in \Theta_\sigma} D_2(d^{sa}_{\boldsymbol{\theta}}||d^{sa}_{\boldsymbol{\theta}'}) \leq \sigma.$$

We call the task of finding a σ-compression of Θ in $\mathcal{M}$ the *policy space compression* problem. Notably, for some $\mathcal{M}, \Theta, \sigma$, a σ-compression of Θ in $\mathcal{M}$ might not exist, as all the policies $\boldsymbol{\theta} \in \Theta$ might induce relevant state-action distributions. Otherwise, we say that the compression is *feasible*. In this case, given $\mathcal{M}$ and Θ, we would like to extract the smallest set of policies Θ' that is a σ-compression of Θ in $\mathcal{M}$, and then keep this reduced policy space to address any RL task one can define over $\mathcal{M}$. Let $\Omega_\Theta := \{d^{sa}_{\boldsymbol{\theta}} \mid \forall \boldsymbol{\theta} \in \Theta\}$ be the set of state-action distributions induced by the policy space Θ, the compression problem can be formulated as a typical *set cover problem*, i.e.,

$$\begin{aligned} \text{minimize} \quad & \sum_{\omega \in \Omega_\Theta} x_\omega \\ \text{subject to} \quad & \sum_{\omega : D_2(\upsilon||\omega) \leq \sigma} x_\omega \geq 1, \quad \forall \upsilon \in \Omega_\Theta \\ & x_\omega \in \{0,1\}, \quad \forall \omega \in \Omega_\Theta \end{aligned} \tag{1}$$

where the positive integers x_ω denote the state-action distributions that are active in the covering, and the corresponding σ-compression of Θ in $\mathcal{M}$ can be retrieved as $\Theta_\sigma = \{\boldsymbol{\theta} \in \Theta \mid d^{sa}_{\boldsymbol{\theta}} = \omega \ \wedge \ x_\omega = 1\}$. Unfortunately, the problem (1) is known to be NP-hard (Feige, 1998), and even building an instance is far-fetched when Ω_Θ is an infinite set. Two aspects arguably make this problem extremely hard: On the one hand, we are looking for an efficient solution in the number of active state-action distributions, secondly, we are covering the set Ω_Θ all at once rather than incrementally. Instead of considering common relaxations of (1) (Johnson, 1974; Lovász, 1975), which would not strictly meet the requirements of Definition 3.1 (Feige, 1998), in the next section we build on these insights to reformulate the policy space compression problem in a tractable way.

4. A Game-Theoretic Reformulation

Due to its inherent hardness, we aim to find a tractable reformulation of the policy space compression problem (1) whose solution is a valid σ-compression of Θ in $\mathcal{M}$. To this end, let us consider a game-theoretic perspective to the set cover problem. In this perspective, a player distributes a set of K policies $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K) \in \Theta^K$ with the intention of covering the set of state-action distributions Ω_Θ. A second player tries to find a policy $\boldsymbol{\mu} \in \Theta$ that is not well covered by $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$, i.e., a policy that maximizes the Rényi divergence between its state-action distribution and the one of the closest $\boldsymbol{\theta}_k \in (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$. The former player moves first, and we call it a *leader*. The latter player makes his move in response to the other player, and it is then called a *follower*. The two-player, zero-sum, sequential game that we have informally described can be represented as the optimization problem

$$\min_{\boldsymbol{\theta} \in \Theta^K} \max_{\boldsymbol{\mu} \in \Theta} \; f(\boldsymbol{\theta}, \boldsymbol{\mu}), \tag{2}$$

$$f(\boldsymbol{\theta}, \boldsymbol{\mu}) := \min_{k \in [K]} D_2(d^{sa}_{\boldsymbol{\mu}} || d^{sa}_{\boldsymbol{\theta}_k}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$ and $[K] = \{0, \dots, K\}$. It is straightforward to see that if the σ-compression is feasible for Θ in $\mathcal{M}$ and K is large enough, then any optimal leader's strategy for the game (2), i.e., $\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta} \in \Theta^K} \min_{\boldsymbol{\mu} \in \Theta} f(\boldsymbol{\theta}, \boldsymbol{\mu})$, is a σ-compression of Θ in $\mathcal{M}$. Unfortunately, $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ is a non-convex non-concave function, and finding a globally optimal strategy for the game (2) is still a NP-hard problem. However, we do not actually need to find a globally optimal strategy for the leader, as any $\boldsymbol{\theta} \in \Theta^K$ such that $\min_{\boldsymbol{\mu} \in \Theta} f(\boldsymbol{\theta}, \boldsymbol{\mu}) \leq \sigma$ would be a valid σ-compression of Θ. Thus, we might instead target a locally optimal strategy for (2), which is a stationary point of f and it is both a local maximum w.r.t. $\boldsymbol{\theta}$ and a local minimum w.r.t. $\boldsymbol{\mu}$. In the next statement, we formalize this solution concept as a Differential Stackelberg Equilibrium (DSE) (Fiez et al., 2020).

Definition 4.1 (Differential Stackelberg (Fiez et al., 2020))**.** *The joint strategy* $(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) \in \Theta^{K+1}$ *in which* $\boldsymbol{\theta}^*_k \in \arg\min_{k \in [K]} (d^{sa}_{\boldsymbol{\mu}^*} || d^{sa}_{\boldsymbol{\theta}^*_k})$ *is a differential Stackelberg equilibrium of the game* (2) *if it holds* $\nabla_{\boldsymbol{\theta}^*_k} f(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) = 0, \nabla_{\boldsymbol{\mu}^*} f(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) = 0, |\nabla_{\boldsymbol{\theta}^*_k} \nabla^\top_{\boldsymbol{\theta}^*_k} f(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*)| > 0,$ *and* $|\nabla_{\boldsymbol{\mu}^*} \nabla^\top_{\boldsymbol{\mu}^*} f(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*)| < 0.$ [1]

Luckily, several recent works have established a favorable complexity for the problem of finding a DSE (Jin et al., 2020b; Fiez et al., 2020; Fiez & Ratliff, 2020) in a sequential game. Especially, Jin et al. (Jin et al., 2020b) showed that a basic first-order method, i.e., Gradient Descent Ascent (GDA), with an infinite time-scale separation between the leader's and follower's updates is guaranteed to converge to a DSE under mild conditions. This result might be surprising, as we started with a fundamentally hard problem (1) and ended up with a way easier formulation (2) that we can address with a common methodology, without making any strong assumption on the structure of the problem. However, we still have to deal with two crucial issues to solve the policy space compression problem through the game-theoretic formulation. On the one hand, it is not enough to look at the value $f(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*)$ attained by a DSE $(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*)$ to guarantee that $\boldsymbol{\theta}$ is a σ-compression of Θ, as we should check that $\max_{\boldsymbol{\mu} \in \Theta} f(\boldsymbol{\theta}^*, \boldsymbol{\mu}) \leq \sigma$, where $\boldsymbol{\mu}$ is a global maximizer. On the other hand, it is not clear how to set a convenient value of K beforehand. In the next section, we present a first-order method that addresses these two issues by finding a DSE of iteratively larger instances of the game (2) (which we will henceforth call the *cover game*) until a conservative approximation of the global condition $\max_{\boldsymbol{\mu} \in \Theta} f(\boldsymbol{\theta}^*, \boldsymbol{\mu}) \leq \sigma$ is finally met.

5. An Algorithm to Solve the Policy Space Compression

Optimization problems of the kind of (2) are typically addressed with a GDA procedure, in which the leader's parameters ($\boldsymbol{\theta}$) and the follower's parameters ($\boldsymbol{\mu}$) are updated iteratively according to

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \boldsymbol{\mu}), \qquad \boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \beta \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\theta}, \boldsymbol{\mu}),$$

where $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \boldsymbol{\mu})$ and $\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\theta}, \boldsymbol{\mu})$ are the respective gradients of the joint objective function. Especially, if we consider a sufficiently large time-scale separation $\tau := \beta / \alpha$, we are guaranteed to converge to a DSE of the game (2) (Jin et al., 2020b; Fiez & Ratliff, 2020). In this case, we can consider $\tau = \infty$, which means we update the follower's parameters until a stationary point is reached, i.e., $\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\theta}, \boldsymbol{\mu}) = 0$, before updating the leader's parameters. However, to instantiate the cover game, we still need to specify the number K of leader-controlled policies $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$. A straightforward solution is to start with a small number of policies first, say $K = 1$, then retrieve a DSE $(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*)$ via GDA for a cover-game instance with K policies, and finally check if the resulting leader's strategy $\boldsymbol{\theta}^*$ meets the global requirement $\max_{\boldsymbol{\mu} \in \Theta} f(\boldsymbol{\theta}^*, \boldsymbol{\mu}) \leq \sigma$. If the answer is positive, the policy space compression problem is solved, and $\boldsymbol{\theta}^*$ is a σ-compression of Θ in $\mathcal{M}$. Otherwise, we increment K and we repeat the process to see if we can solve the problem with more policies in $\boldsymbol{\theta}$. If the policy space compression problem is feasible, with this simple procedure we are guaranteed to

[1] Let $f(\boldsymbol{x})$ be a function of $\boldsymbol{x} \in \mathbb{R}^m$, we denote its gradient vector as $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$, its Hessian matrix as $\nabla_{\boldsymbol{x}} \nabla^\top_{\boldsymbol{x}} f(\boldsymbol{x})$, and the determinant of its Hessian matrix as $|\nabla_{\boldsymbol{x}} \nabla^\top_{\boldsymbol{x}} f(\boldsymbol{x})|$.

Algorithm 1 PSCA

```
Input: CMP $\mathcal{M}$, policy space $\Theta$, constant $\sigma$
initialize $K = 0$ and the cover guarantee $\overline{\mathcal{Z}}_{\boldsymbol{\theta}} = \infty$
while $\overline{\mathcal{Z}}_{\boldsymbol{\theta}} > \sigma$ do
  $K \leftarrow K + 1$
  initialize the leader $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K) \in \Theta^K$
  for epoch = $1, 2, \dots,$ until convergence do
    compute the best response $\boldsymbol{\mu}_{br}$ to $\boldsymbol{\theta}$
    identify the active leader's component w.r.t. $\boldsymbol{\mu}_{br}$
    update the leader $\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}_k} f(\boldsymbol{\theta}, \boldsymbol{\mu}_{br})$
  end for
  compute the cover guarantee $\overline{\mathcal{Z}}_{\boldsymbol{\theta}}$ with (6)
end while
Output: return $\boldsymbol{\theta}$, i.e., a $\sigma$-compression of $\Theta$ in $\mathcal{M}$
```

get a valid σ-compression eventually. We call this method the *Policy Space Compression Algorithm* (PSCA) and we report the pseudocode in Algorithm 1. In the following sections, we describe in details how the optimization of the follower's parameters (Section 5.1) and the leader's parameters (Section 5.2) are carried out in an adaptation of the GDA method to the specific setting of the cover game. In Section 5.3, we discuss how to verify the global requirement $\max_{\boldsymbol{\mu} \in \Theta} f(\boldsymbol{\theta}^*, \boldsymbol{\mu}) \leq \sigma$ without actually having to find a globally optimal follower's strategy, but instead optimizing a surrogate objective through a tractable linear program.

5.1. Optimizing the Follower's Parameters

In principle, we would like to compute the gradient $\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\theta}, \boldsymbol{\mu})$ to perform the update $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \beta \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\theta}, \boldsymbol{\mu})$ as in a common GDA procedure. Unfortunately, the objective function $f(\boldsymbol{\theta}, \boldsymbol{\mu}) = \min_{k \in [K]} D_2(d^{sa}_{\boldsymbol{\mu}} || d^{sa}_{\boldsymbol{\theta}_k})$ is not differentiable due to the minimum over the K components of $\boldsymbol{\theta}$. However, only the leader's component $\boldsymbol{\theta}_k$ that attains the minimum of f is actually relevant for the follower's update, as the other $K-1$ components do not affect the value of the objective. Thus, we call $\boldsymbol{\theta}_k \in \arg\min_{\boldsymbol{\theta}_i \in \boldsymbol{\theta}} D_2(d^{sa}_{\boldsymbol{\mu}} || d^{sa}_{\boldsymbol{\theta}_k})$ the *active leader's component*. Conveniently, we can update the follower's parameters w.r.t. the gradient $\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\theta}_k, \boldsymbol{\mu})$, which is differentiable w.r.t. $\boldsymbol{\mu}$. The following proposition provides the formula for this gradient.

Proposition 5.1 (Follower's Gradient)**.** *Let $(\boldsymbol{\theta}, \boldsymbol{\mu}) \in \Theta^K$, the gradient of $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ w.r.t. $\boldsymbol{\mu}$ is given by*

$$\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\theta}, \boldsymbol{\mu}) =$$

$$2 \mathop{\mathbb{E}}_{(s,a) \sim d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \left[\left(\frac{d^{sa}_{\boldsymbol{\mu}}(s,a)}{d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \right)^2 \nabla_{\boldsymbol{\mu}} \log d^{sa}_{\boldsymbol{\mu}}(s,a) \right], \quad (3)$$

where $\boldsymbol{\theta}_k$ is the leader's component such that $\boldsymbol{\theta}_k \in \arg\min_{\boldsymbol{\theta}_i \in \boldsymbol{\theta}} D_2(d^{sa}_{\boldsymbol{\mu}} || d^{sa}_{\boldsymbol{\theta}_i})$.

To perform a full optimization of the follower's parameters, we just need to repeatedly apply the gradient ascent update with the gradient $\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\theta}, \boldsymbol{\mu})$ computed as in (3). Under mild conditions on the learning rate (Robbins & Monro, 1951), this process is guaranteed to converge to a stationary point such that $\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\theta}, \boldsymbol{\mu}) = 0$. We call the follower's parameters $\boldsymbol{\mu}$ at this stationary point the *best response* to the leader's parameter $\boldsymbol{\theta}$, and we denote it as $\boldsymbol{\mu}_{br}$.

5.2. Optimizing the Leader's Parameters

Whenever the follower converges at the best response $\boldsymbol{\mu}_{br}$ to the current leader's parameters, we would like to make an update to $\boldsymbol{\theta}$ in the direction of the gradient $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \boldsymbol{\mu})$, i.e., $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \boldsymbol{\mu})$. Just as before, we can pre-compute the active leader's component $\boldsymbol{\theta}_k \in \arg\min_{\boldsymbol{\theta}_i \in \boldsymbol{\theta}} D_2(d^{sa}_{\boldsymbol{\mu}} || d^{sa}_{\boldsymbol{\theta}_i})$ to make an update to $\boldsymbol{\theta}_k$ in the direction of the gradient $\nabla_{\boldsymbol{\theta}_k} f(\boldsymbol{\theta}_k, \boldsymbol{\mu})$, which is differentiable in $\boldsymbol{\theta}_k$. Indeed, an update to any other leader's component would not have a meaningful impact on the value of the objective, whereas updating $\boldsymbol{\theta}_k$ with a sufficiently small learning rate α is guaranteed to decrease $f(\boldsymbol{\theta}, \boldsymbol{\mu})$, possibly forcing the follower to change its best response in the next epoch. The following proposition provides the formula for the gradient.

Proposition 5.2 (Leader's Gradient)**.** *Let $(\boldsymbol{\theta}, \boldsymbol{\mu}) \in \Theta^K$, the gradient of $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ w.r.t. $\boldsymbol{\theta}_k$ is given by*

$$\nabla_{\boldsymbol{\theta}_k} f(\boldsymbol{\theta}, \boldsymbol{\mu}) =$$

$$- \mathop{\mathbb{E}}_{(s,a) \sim d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \left[\left(\frac{d^{sa}_{\boldsymbol{\mu}}(s,a)}{d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \right)^2 \nabla_{\boldsymbol{\mu}} \log d^{sa}_{\boldsymbol{\theta}_k}(s,a) \right]. \quad (4)$$

5.3. Assessing the Global Value of the Leader's Parameters

The last missing piece of the PSCA algorithm requires verifying that the leader's strategy in the DSE $(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*)$ obtained from the GDA procedure is actually a σ-compression of Θ in $\mathcal{M}$. In principle, we should verify that $\min_{k \in [K]} D_2(\boldsymbol{\theta}^*, \boldsymbol{\mu}) \leq \sigma$ for any $\boldsymbol{\mu} \in \Theta$, which is equivalent to controlling if $\max_{\boldsymbol{\mu} \in \Theta} f(\boldsymbol{\theta}^*, \boldsymbol{\mu}) \leq \sigma$. Unfortunately, the follower's strategy $\boldsymbol{\mu}^*$ is only locally optimal. Thus, checking $f(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) \leq \sigma$ is not sufficient, as the globally optimal follower's strategy might attain a greater value of f than $\boldsymbol{\mu}^*$. Instead, we should check $\mathcal{Z}_{\boldsymbol{\theta}^*} \leq \sigma$, where $\mathcal{Z}_{\boldsymbol{\theta}^*}$ is given by

$$\mathcal{Z}_{\boldsymbol{\theta}^*} = \max_{\omega \in \Omega_\Theta} \min_{k \in [K]} \int_{\mathcal{SA}} \big(\omega(s,a)\big)^2 \big(d^{sa}_{\boldsymbol{\theta}^*_k}(s,a)\big)^{-1} \, \mathrm{d}s \, \mathrm{d}a, \quad (5)$$

which can be written as a quadratically constrained quadratic program (see Appendix B.1). It might come as no surprise that solving this problem is NP-hard. Indeed, this is equivalent to the problem (2) with a fixed leader's strategy $\boldsymbol{\theta}^*$, but the objective $f(\boldsymbol{\theta}^*, \boldsymbol{\mu})$ is still non-concave w.r.t. $\boldsymbol{\mu}$. Luckily, we can reformulate this NP-hard problem in the surrogate linear program (see Appendix B.2):

$$\overline{\mathcal{Z}}_{\boldsymbol{\theta}^*} = \max_{\omega \in \Omega_\Theta} \min_{k \in [K]} \int_{\mathcal{SA}} \omega(s,a) \big(d^{sa}_{\boldsymbol{\theta}^*_k}(s,a) \big)^{-\frac{1}{2}} \, \mathrm{d}s \, \mathrm{d}a, \quad (6)$$

where the value $(\overline{\mathcal{Z}}_{\boldsymbol{\theta}^*})^2$ is a conservative approximation of $\mathcal{Z}_{\boldsymbol{\theta}^*}$, as stated in the following theorem:

Theorem 5.3. *The value $(\overline{\mathcal{Z}}_{\boldsymbol{\theta}^*})^2$ is an upper bound to the value $\mathcal{Z}_{\boldsymbol{\theta}^*}$, i.e., $(\overline{\mathcal{Z}}_{\boldsymbol{\theta}^*})^2 \geq \mathcal{Z}_{\boldsymbol{\theta}^*}, \forall \boldsymbol{\theta}^* \in \Theta^K$.*

6. Theoretical Guarantees of RL with a Compressed Policy Space

In the previous sections, we have motivated the pursuit of a compression of the original policy space Θ in the CMP $\mathcal{M}$ as a way to improve the computational and sample efficiency of solving RL tasks defined upon $\mathcal{M}$, and we have presented a viable methodology for extracting such a compression Θ_σ. Since this compression procedure induces a loss, albeit bounded, in the expressive power of the policy space, it is worth investigating the performance guarantees that we have when addressing RL tasks with Θ_σ. We first analyze *policy evaluation* (Section 6.1) and then *policy optimization* (Section 6.2). The reported theoretical results mostly combine techniques from (Metelli et al., 2018; Papini et al., 2019).

6.1. Policy Evaluation

In policy evaluation (Sutton & Barto, 2018), we aim to estimate the performance $J(\boldsymbol{\theta})$ of a target policy $\boldsymbol{\theta} \in \Theta$ through sampled interactions with an MDP $\mathcal{M}^{\mathcal{R}}$. In our case, we can only draw samples with the policies in Θ_σ, and we have to provide an off-policy estimate of $J(\boldsymbol{\theta})$ via importance sampling. Since for any target policy $\boldsymbol{\theta}$ we are guaranteed to have a sampling policy $\boldsymbol{\theta}' \in \Theta_\sigma$ such that $D_2(d^{sa}_{\boldsymbol{\theta}} || d^{sa}_{\boldsymbol{\theta}'}) \leq \sigma$, by choosing a convenient sampling policy in Θ_σ, we can enjoy the following guarantee on the error we would make when evaluating any target policy $\boldsymbol{\theta} \in \Theta$ in any MDP $\mathcal{M}^{\mathcal{R}}$ one can build upon $\mathcal{M}$.

Theorem 6.1 (Evaluation Error)**.** *Let Θ_σ be a σ-compression of Θ in $\mathcal{M}$, let $\mathcal{R}$ be a reward function for $\mathcal{M}$ uniformly bounded by $\mathrm{R}_{\max}$, let $\boldsymbol{\theta} \in \Theta$ be a target policy, and let $\delta \in (0,1)$ be a confidence. There exists $\boldsymbol{\theta}' \in \Theta_\sigma$ such that, given N samples from $\boldsymbol{\theta}'$, the error of the importance sampling evaluation of $J(\boldsymbol{\theta})$ in $\mathcal{M}^{\mathcal{R}}$, i.e., $\widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}') = \frac{1}{(1-\gamma)N} \sum_{n=1}^{N} w_{\boldsymbol{\theta}/\boldsymbol{\theta}'}(s_n, a_n) \mathcal{R}(s_n, a_n)$, is upper bounded as $|J(\boldsymbol{\theta}) - \widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')| \leq \frac{\mathrm{R}_{\max}}{1-\gamma} \sqrt{\sigma / \delta N}$ with probability at least $1 - \delta$.*

Notably, given a budget of samples N, a confidence δ, and a requirement on the evaluation error beforehand, we could select a proper σ to build a σ-compression that meets the requirement in any policy evaluation task. However, choosing a sampling policy $\boldsymbol{\theta}' \in \Theta_\sigma$ that is best suited for a given task might be non-trivial. Thus, one can instead take a batch of N_k samples with each policy in Θ_σ, and then perform the policy evaluation via Multiple Importance Sampling (MIS) (Owen, 2013; Papini et al., 2019).

Corollary 6.2. *Let Θ_σ be a σ-compression of Θ in $\mathcal{M}$ such that $|\Theta_\sigma| = K$, let $\mathcal{R}$ be a reward function for $\mathcal{M}$ uniformly bounded by $\mathrm{R}_{\max}$, let $\boldsymbol{\theta} \in \Theta$ be a target policy, and let $\delta \in (0,1)$ be a confidence. Given N_k samples from each $\boldsymbol{\theta}_k \in \Theta_\sigma$, the error of the multiple importance sampling evaluation of $J(\boldsymbol{\theta})$ in $\mathcal{M}^{\mathcal{R}}$, i.e.,*

$$\widehat{J}_{MIS}(\boldsymbol{\theta}/\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K) = \frac{1}{(1-\gamma)} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \frac{d^{sa}_{\boldsymbol{\theta}}(s_{n,k}, a_{n,k})}{\sum_{j=1}^{K} N_j d^{sa}_{\boldsymbol{\theta}_j}(s_{n,k}, a_{n,k})} \mathcal{R}(s_{n,k}, a_{n,k}),$$

is upper bounded as $|J(\boldsymbol{\theta}) - \widehat{J}_{MIS}(\boldsymbol{\theta}/\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)| \leq \frac{\mathrm{R}_{\max}}{1-\gamma} \sqrt{D_2(d^{sa}_{\boldsymbol{\theta}} || \Phi) / \delta N}$ with probability at least $1 - \delta$, where $N = \sum_{k=1}^{K} N_k$ is a number of samples and $\Phi = \sum_{k=1}^{K} \frac{N_k}{N} d^{sa}_{\boldsymbol{\theta}_k}$ is a finite mixture.

Thanks to the result in (Metelli et al., 2020b, Theorem 1), in tabular MDPs the evaluation error of the MIS estimator is guaranteed to be lower than the one of the IS estimator of Theorem 6.1 (as long as $N_k \geq N$, where N is the number of samples considered by the IS estimator).

6.2. Policy Optimization

In policy optimization (see Section 2.2), we seek for the policy $\boldsymbol{\theta}$ that maximizes $J(\boldsymbol{\theta})$ within a parametric policy space. In principle, we could look for the policy that maximizes the performance within the σ-compression Θ_σ, which can be found efficiently with the OPTIMIST algorithm (Papini et al., 2019). Especially, in this setting OPTIMIST yields constant regret for tabular MDPs (Metelli et al., 2020a), as the set Θ_σ is finite and it is composed of stochastic policies such that $\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_\sigma, D_2(d^{sa}_{\boldsymbol{\theta}} || d^{sa}_{\boldsymbol{\theta}'}) < \infty$. However, this optimal policy within Θ_σ might be sub-optimal w.r.t. the optimal policy within the original policy space Θ. We can still upper bound this sub-optimality, as reported in the following theorem.

Theorem 6.3 (Policy Optimization in Θ_σ)**.** *Let Θ_σ be a σ-compression of Θ in $\mathcal{M}$, and let $\mathcal{R}$ be a reward function for $\mathcal{M}$ uniformly bounded by $\mathrm{R}_{\max}$. The policy $\boldsymbol{\theta}^*_\sigma \in \arg\max_{\boldsymbol{\theta} \in \Theta_\sigma} J(\boldsymbol{\theta})$ is ϵ-optimal for the MDP $\mathcal{M}^{\mathcal{R}}$, where $\epsilon := |\max_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) - J(\boldsymbol{\theta}^*_\sigma)| \leq \frac{\mathrm{R}_{\max}}{1-\gamma} \sqrt{\log \sigma}$.*

Notably, the latter guarantee does not involve any estimation, and the policy $\boldsymbol{\theta}^*$ can be obtained in a finite number of interactions. Nonetheless, one can shrink even more the sub-optimality ϵ, and without deteriorating the sample complexity, by coupling the OPTIMIST algorithm with an additional offline optimization procedure. The idea is to return the policy $\boldsymbol{\theta} \in \Theta$ that maximizes the importance sampling evaluation obtained with the samples from the policies in Θ_σ.

Theorem 6.4 (Off-Policy Optimization in Θ)**.** *Let Θ_σ be a σ-compression of Θ in $\mathcal{M}$ such that $|\Theta_\sigma| = K$, let $\mathcal{R}$ be a reward function for $\mathcal{M}$ uniformly bounded by $\mathrm{R}_{\max}$, and let $\delta \in (0, 1)$ be a confidence. Given N_k samples from each $\boldsymbol{\theta}_k \in \Theta_\sigma$, we can recover an ϵ-optimal policy for $\mathcal{M}^{\mathcal{R}}$ as*

$$\left(\_ \, , \, \boldsymbol{\theta}^*_{IS} \right) \in \operatorname*{arg\,max}_{\boldsymbol{\theta}_k \in \Theta_\sigma, \boldsymbol{\theta} \in \Theta : D_2(d^{sa}_{\boldsymbol{\theta}} || d^{sa}_{\boldsymbol{\theta}_k})} \frac{1}{(1-\gamma) N_k} \sum_{n=1}^{N_k} w_{\boldsymbol{\theta}/\boldsymbol{\theta}_k}(s_n, a_n) \mathcal{R}(s_n, a_n), \quad (7)$$

such that $\epsilon := \left| \max_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) - J(\boldsymbol{\theta}^*_{IS}) \right| \leq \frac{\mathrm{R}_{\max}}{1-\gamma} \sqrt{2\sigma / N_k \delta}$ with probability at least $1 - \delta$.

Although, contrary to the guarantee in Theorem 6.3, ϵ vanishes with the number of samples in the latter result, solving problem (7) is non-trivial in general, as the policy space Θ is often infinite.

7. Numerical Validation

In this section, we provide a brief numerical validation of the policy space compression problem (Section 7.1) and how it benefits RL (Section 7.2, 7.3). To the purpose of the analysis, we consider the *River Swim* domain (Strehl & Littman, 2008), in which an agent navigates a chain of six states by taking one of two actions: either *swim up*, to move upstream towards the upper states, or *swim down*, to go downstream back to the lower states. In Appendix C, we report further details on the experimental settings, along with some additional results in a *Grid World* environment. We leave as future work a more extensive experimental evaluation of the policy space compression problem beyond toy domains.

7.1. Policy Space Compression

In the River Swim, we consider the policy space $\Theta \subseteq \mathbb{R}^{|\mathcal{S}| \times (|\mathcal{A}|-1)}$ of the *softmax* policies $\pi_{\boldsymbol{\theta}}(a|s) = \exp(\theta_{sa}) / \sum_{j \in \mathcal{A}} \exp(\theta_{sj})$, and we seek for a compression Θ_σ with the requirement $\sigma = 10$, such that Θ_σ is a valid σ-compression if $\min_{\boldsymbol{\theta} \in \Theta_\sigma} \max_{\boldsymbol{\mu} \in \Theta} f(\boldsymbol{\theta}, \boldsymbol{\mu}) \leq 10$. In Figure 1a, we report the values of $\mathcal{Z} = \max_{\boldsymbol{\mu} \in \Theta} f(\boldsymbol{\theta}, \boldsymbol{\mu})$ (5) and its upper bound $\overline{\mathcal{Z}} \geq \mathcal{Z}$ (6). Especially, we can see that PSCA effectively founds a valid σ-compression Θ_σ of just $K = 3$ policies (Figure 1a, left), and that the values of $\mathcal{Z}$ and $\overline{\mathcal{Z}}$ smoothly decreases during the GDA procedure for a fixed number of policies (Figure 1a, right). Notably, $K = 2$ policies are actually sufficient to meet the σ requirement in this setting. However, PSCA cannot access $\mathcal{Z}$ but its conservative approximation $\overline{\mathcal{Z}}$, and thus stops whenever $\overline{\mathcal{Z}} \leq \sigma$. In Appendix C, we report an illustration of the obtained policies $\boldsymbol{\theta}_k \in \Theta_\sigma$.

7.2. Policy Evaluation with a Compressed Policy Space

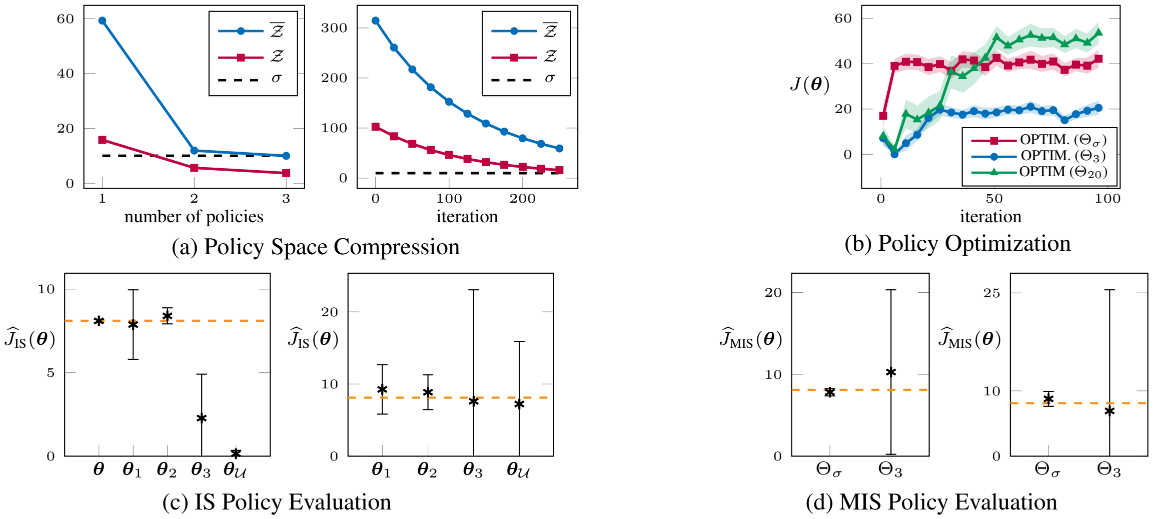
We now aim to show that the obtained σ-compression Θ_σ can be employed with benefit in the most challenging policy evaluation task one can define in the River Swim, which is the off-policy evaluation of an ϵ-greedy policy $\boldsymbol{\theta}$ for the reward function $\mathcal{R}$ that assigns $\mathrm{R}_{\max} = 100$ for taking the action *swim up* in the final state. In Figure 1c, we show that sampling with the policies $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_\sigma$ lead to an IS off-policy evaluation that is comparable to the exact $J(\boldsymbol{\theta})$ (dashed line) and its on-policy estimate $(\boldsymbol{\theta})$. Even by sampling from a uniform mixture of the policies in Θ_σ, the performance of the MIS evaluation is significantly better than the one obtained by a uniform mixture of three random policies (Θ_3), as reported in Figure 1d. For both the IS and the MIS regime, we provide the empirical evaluations (on the left) and the hindsight evaluations (right) obtained with the exact values of the importance weights $w_{\boldsymbol{\theta}/\boldsymbol{\theta}'}$ and the confidence bounds of the Theorem 6.1, 6.2 respectively.

7.3. Compression for Policy Optimization

Finally, we show that the compression Θ_σ allows for efficient policy optimization. We consider the same reward function $\mathcal{R}$ of the previous section, and the OPTIMIST (Papini et al., 2019) algorithm equipped with Θ_σ, or a uniform discretization of the original policy space Θ with either three policies (Θ_3) or twenty policies (Θ_{20}). In Figure 1b, we show that OPTIMIST with Θ_σ swiftly converges (less than five iterations) to the optimal policy within the space. Instead, the policy space Θ_3 leads to a huge sub-optimality in the final performance, and OPTIMIST with Θ_{20} is way slower to converge to the optimal policy within the space. These results are a testament of the ability of PSCA to incorporate the peculiar structure of the domain in a small set of representative policies Θ_σ, and to allows for a remarkable balance between sample efficiency and sub-optimality in subsequent policy optimization.

8. Conclusions

In this paper, we considered the problem of compressing an infinite parametric policy space into a finite set of representative policies for a given environment. First, we provided a formal definition of the problem, and we highlighted its inherent hardness. Then, we proposed a tractable game-theoretic reformulation, for which a locally optimal solution can be efficiently found through an iterative GDA procedure. Finally, we provided a theoretical characterization of the guarantees that the compression brings to subsequent RL



(a) Policy Space Compression

(b) Policy Optimization

(c) IS Policy Evaluation

(d) MIS Policy Evaluation

Figure 1. Set of experiments in the *River Swim* domain. **(a)** The value of the compression guarantee $\mathcal{Z}$, its upper bound $\overline{\mathcal{Z}}$, and the requirement σ as a function of the number of policies K (left) and as a function of the iterations with $K = 1$ (right) obtained with PSCA. **(b)** The average return $J(\boldsymbol{\theta})$ obtained by OPTIMIST with the σ-compression Θ_σ (3 policies), a 3-policies discretization Θ_3, and a 20-policies discretization Θ_{20} (95% c.i. over 50 runs). **(c,d)** IS and MIS evaluation of $J(\boldsymbol{\theta})$ by taking samples with $\boldsymbol{\theta}$, $\boldsymbol{\theta}_k \in \Theta_\sigma$, a uniform policy $\boldsymbol{\theta}_{\mathcal{U}}$, the mixture Θ_σ, or a mixture of 3 random policies Θ_3. We provide both the empirical (left, 95% c.i. over 50 runs) and the hindsight (right) values.

tasks, and a numerical validation of the proposed approach.

Future works might target the compression problem from interactions with an *unknown* environment, to pave the way for scalable policy space compression. Especially, such an extension would require sample-based estimates of the gradients (3), (4), and the global guarantee (6). Whereas estimating the gradients of state-action distributions is not an easy feat, previous works provide useful inspiration (Morimura et al., 2010; Schroecker & Isbell, 2017; Schroecker et al., 2018). Other interesting future directions include an extension of the policy space compression problem to the parameter-based perspective (Sehnke et al., 2008; Metelli et al., 2018; Papini et al., 2019), and the development of policy optimization algorithms that are tailored to exploit a compression of the policy space.

References

Cochran, W. G. *Sampling techniques*. John Wiley & Sons, 2007.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, 2010.

Deisenroth, M. P., Neumann, G., Peters, J., et al. A survey on policy search for robotics. *Foundations and trends in Robotics*, 2(1-2):388–403, 2013.

Feige, U. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 1998.

Fiez, T. and Ratliff, L. Gradient descent-ascent provably converges to strict local minmax equilibria with a finite timescale separation. *arXiv preprint arXiv:2009.14820*, 2020.

Fiez, T., Chasnov, B., and Ratliff, L. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *Proceedings of the International Conference on Machine Learning*, 2020.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. Provably efficient maximum entropy exploration. In *Proceedings of the International Conference on Machine Learning*, 2019.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2020a.

Jin, C., Netrapalli, P., and Jordan, M. What is local optimality in nonconvex-nonconcave minimax optimization? In *Proceedings of the International Conference on Machine Learning*, 2020b.

Johnson, D. S. Approximation algorithms for combinatorial problems. *Journal of computer and system sciences*, 1974.

Lovász, L. On the ratio of optimal integral and fractional covers. *Discrete mathematics*, 1975.

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

Metelli, A. M., Papini, M., D'Oro, P., and Restelli, M. Policy optimization as online learning with mediator feedback. *arXiv preprint arXiv:2012.08225*, 2020a.

Metelli, A. M., Papini, M., Montali, N., and Restelli, M. Importance sampling techniques for policy optimization. *Journal of Machine Learning Research*, 2020b.

Morimura, T., Uchibe, E., Yoshimoto, J., Peters, J., and Doya, K. Derivatives of logarithmic stationary distributions for policy gradient reinforcement learning. *Neural computation*, 2010.

Owen, A. B. *Monte Carlo theory, methods and examples*. 2013.

Papini, M., Metelli, A. M., Lupo, L., and Restelli, M. Optimistic policy optimization via multiple importance sampling. In *Proceedings of the International Conference on Machine Learning*, 2019.

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

Rényi, A. et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, 1951.

Schroecker, Y. and Isbell, C. L. State aware imitation learning. *Advances in Neural Information Processing Systems*, 2017.

Schroecker, Y., Vecerik, M., and Scholz, J. Generative predecessor models for sample-efficient imitation learning. In *International Conference on Learning Representations*, 2018.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *Proceedings of the International Conference on Machine Learning*, 2015.

Sehnke, F., Osendorfer, C., Rückstieß, T., Graves, A., Peters, J., and Schmidhuber, J. Policy gradients with parameter-based exploration for control. In *International Conference on Artificial Neural Networks*. Springer, 2008.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *Proceedings of the International Conference on Machine Learning*, 2014.

Sobel, M. J. The variance of discounted markov decision processes. *Journal of Applied Probability*, 1982.

Strehl, A. L. and Littman, M. L. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 2008.

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

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

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, 2019.

A. Proofs

A.1. Proofs of Section 5

Proposition 5.1 (Follower's Gradient)**.** *Let $(\boldsymbol{\theta}, \boldsymbol{\mu}) \in \Theta^K$, the gradient of $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ w.r.t. $\boldsymbol{\mu}$ is given by*

$$\begin{aligned}&\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\theta}, \boldsymbol{\mu}) = \\ &2 \mathop{\mathbb{E}}_{(s,a)\sim d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \left[\left(\frac{d^{sa}_{\boldsymbol{\mu}}(s,a)}{d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \right)^2 \nabla_{\boldsymbol{\mu}} \log d^{sa}_{\boldsymbol{\mu}}(s,a) \right],\end{aligned} \tag{3}$$

where $\boldsymbol{\theta}_k$ is the leader's component such that $\boldsymbol{\theta}_k \in \arg\min_{\boldsymbol{\theta}_i \in \boldsymbol{\theta}} D_2(d^{sa}_{\boldsymbol{\mu}} || d^{sa}_{\boldsymbol{\theta}_i})$.

Proof. Let $\boldsymbol{\theta}_k$ be the active leader's component, i.e., $\boldsymbol{\theta}_k \in \arg\min_{\boldsymbol{\theta}_i \in \boldsymbol{\theta}} D_2(d^{sa}_{\boldsymbol{\mu}} || d^{sa}_{\boldsymbol{\theta}_i})$. We can compute the gradient of the objective $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ w.r.t. $\boldsymbol{\mu}$ as

$$\begin{aligned}\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\theta}, \boldsymbol{\mu}) &= \nabla_{\boldsymbol{\mu}} D_2(d^{sa}_{\boldsymbol{\mu}} || d^{sa}_{\boldsymbol{\theta}_k}) \\ &= \nabla_{\boldsymbol{\mu}} \int_{\mathcal{SA}} d^{sa}_{\boldsymbol{\theta}_k}(s,a) \left(\frac{d^{sa}_{\boldsymbol{\mu}}(s,a)}{d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \right)^2 \mathrm{d}s \, \mathrm{d}a \\ &= 2 \int_{\mathcal{SA}} d^{sa}_{\boldsymbol{\theta}_k}(s,a) \left(\frac{d^{sa}_{\boldsymbol{\mu}}(s,a)}{d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \right)^2 \nabla_{\boldsymbol{\mu}} \log d^{sa}_{\boldsymbol{\mu}}(s,a) \, \mathrm{d}s \, \mathrm{d}a.\end{aligned}$$

□

Proposition 5.2 (Leader's Gradient)**.** *Let $(\boldsymbol{\theta}, \boldsymbol{\mu}) \in \Theta^K$, the gradient of $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ w.r.t. $\boldsymbol{\theta}_k$ is given by*

$$\begin{aligned}&\nabla_{\boldsymbol{\theta}_k} f(\boldsymbol{\theta}, \boldsymbol{\mu}) = \\ &- \mathop{\mathbb{E}}_{(s,a)\sim d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \left[\left(\frac{d^{sa}_{\boldsymbol{\mu}}(s,a)}{d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \right)^2 \nabla_{\boldsymbol{\mu}} \log d^{sa}_{\boldsymbol{\theta}_k}(s,a) \right].\end{aligned} \tag{4}$$

Proof. We can compute the gradient of the objective $f(\boldsymbol{\theta}, \boldsymbol{\mu})$ w.r.t. $\boldsymbol{\theta}_k \in \boldsymbol{\theta}$ as

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_k} f(\boldsymbol{\theta}, \boldsymbol{\mu}) &= \nabla_{\boldsymbol{\theta}_k} D_2(d^{sa}_{\boldsymbol{\mu}} || d^{sa}_{\boldsymbol{\theta}_k}) \\ &= \nabla_{\boldsymbol{\theta}_k} \int_{\mathcal{SA}} d^{sa}_{\boldsymbol{\theta}_k}(s,a) \left(\frac{d^{sa}_{\boldsymbol{\mu}}(s,a)}{d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \right)^2 \mathrm{d}s \, \mathrm{d}a \\ &= - \int_{\mathcal{SA}} d^{sa}_{\boldsymbol{\theta}_k}(s,a) \left(\frac{d^{sa}_{\boldsymbol{\mu}}(s,a)}{d^{sa}_{\boldsymbol{\theta}_k}(s,a)} \right)^2 \nabla_{\boldsymbol{\theta}_k} \log d^{sa}_{\boldsymbol{\theta}_k}(s,a) \, \mathrm{d}s \, \mathrm{d}a.\end{aligned}$$

□

Theorem 5.3. *The value $(\overline{\mathcal{Z}}_{\boldsymbol{\theta}^*})^2$ is an upper bound to the value $\mathcal{Z}_{\boldsymbol{\theta}^*}$, i.e., $(\overline{\mathcal{Z}}_{\boldsymbol{\theta}^*})^2 \geq \mathcal{Z}_{\boldsymbol{\theta}^*}, \forall \boldsymbol{\theta}^* \in \Theta^K$.*

Proof. The result is straightforward from

$$\begin{aligned}(\overline{\mathcal{Z}}_{\boldsymbol{\theta}^*})^2 &= \max_{\omega \in \Omega_\Theta} \min_{k \in [K]} \left(\int_{\mathcal{SA}} \omega(s,a) \big(d^{sa}_{\boldsymbol{\theta}^*_k}(s,a) \big)^{-\frac{1}{2}} \mathrm{d}s \, \mathrm{d}a \right)^2 \\ &\geq \max_{\omega \in \Omega_\Theta} \min_{k \in [K]} \int_{\mathcal{SA}} \Big(\omega(s,a) \big(d^{sa}_{\boldsymbol{\theta}^*_k}(s,a) \big)^{-\frac{1}{2}} \Big)^2 \mathrm{d}s \, \mathrm{d}a = \mathcal{Z}_{\boldsymbol{\theta}^*}.\end{aligned}$$

□

A.2. Proofs of Section 6

Lemma A.1 (Variance of the IS Estimator)**.** *Let $\mathcal{M}$ be a CMP, and let $\boldsymbol{\theta} \in \Theta$ be a target policy. Let $\{s_n, a_n\}_{n=1}^N$ be a sample of state-action pairs taken with the policy $\boldsymbol{\theta}'$ in $\mathcal{M}$. Then, the variance of the importance sampling evaluation of $J(\boldsymbol{\theta})$ in $\mathcal{M}$, i.e., $\widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}') = \frac{1}{(1-\gamma)N}\sum_{n=1}^N w_{\boldsymbol{\theta}/\boldsymbol{\theta}'}(s_n, a_n)\mathcal{R}(s_n, a_n)$, can be upper bounded as*

$$\underset{(s,a)\sim d^{sa}_{\boldsymbol{\theta}'}}{\mathbb{V}\mathrm{ar}} \left[\widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')\right] \leq \frac{(\mathrm{R}_{\max})^2 D_2(d^{sa}_{\boldsymbol{\theta}}||d^{sa}_{\boldsymbol{\theta}'})}{(1-\gamma)^2\, N}.$$

Proof. The proof follows the derivation in (Metelli et al., 2018, Lemma 4.1). When considering state-action pairs (as opposed to trajectories in (Metelli et al., 2018)) one should account for the dependency between state-actions in the same trajectory. Here we consider a batch of N samples taken within a single trajectory, in which the dependency vanishes as the CMP mixes to the steady-state, and thus it can be neglected. Especially, we write

$$\begin{aligned}\underset{(s,a)\sim d^{sa}_{\boldsymbol{\theta}'}}{\mathbb{V}\mathrm{ar}} \left[\widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')\right] &\leq \frac{1}{(1-\gamma)^2 N} \underset{(s,a)\sim d^{sa}_{\boldsymbol{\theta}'}}{\mathbb{V}\mathrm{ar}} \left[w_{\boldsymbol{\theta}/\boldsymbol{\theta}'}(s,a)\mathcal{R}(s,a)\right] \\ &\leq \frac{1}{(1-\gamma)^2 N} \underset{(s,a)\sim d^{sa}_{\boldsymbol{\theta}'}}{\mathbb{E}} \left[\left(\frac{d^{sa}_{\boldsymbol{\theta}}(s,a)}{d^{sa}_{\boldsymbol{\theta}'}(s,a)}\mathcal{R}(s,a)\right)^2\right] \\ &\leq \frac{(\mathrm{R}_{\max})^2}{(1-\gamma)^2 N} \underset{(s,a)\sim d^{sa}_{\boldsymbol{\theta}'}}{\mathbb{E}} \left[\left(\frac{d^{sa}_{\boldsymbol{\theta}}(s,a)}{d^{sa}_{\boldsymbol{\theta}'}(s,a)}\right)^2\right] = \frac{(\mathrm{R}_{\max})^2 D_2(d^{sa}_{\boldsymbol{\theta}}||d^{sa}_{\boldsymbol{\theta}'})}{(1-\gamma)^2\, N}.\end{aligned}$$

In episodic settings, one can refine this result to account for dependent data by exploiting the Bellman equation of the variance (see Sobel, 1982; Xie et al., 2019). □

Theorem 6.1 (Evaluation Error)**.** *Let Θ_σ be a σ-compression of Θ in $\mathcal{M}$, let $\mathcal{R}$ be a reward function for $\mathcal{M}$ uniformly bounded by $\mathrm{R}_{\max}$, let $\boldsymbol{\theta} \in \Theta$ be a target policy, and let $\delta \in (0, 1)$ be a confidence. There exists $\boldsymbol{\theta}' \in \Theta_\sigma$ such that, given N samples from $\boldsymbol{\theta}'$, the error of the importance sampling evaluation of $J(\boldsymbol{\theta})$ in $\mathcal{M}^{\mathcal{R}}$, i.e., $\widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}') = \frac{1}{(1-\gamma)N}\sum_{n=1}^N w_{\boldsymbol{\theta}/\boldsymbol{\theta}'}(s_n, a_n)\mathcal{R}(s_n, a_n)$, is upper bounded as $|J(\boldsymbol{\theta}) - \widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')| \leq \frac{\mathrm{R}_{\max}}{1-\gamma}\sqrt{\sigma/\delta N}$ with probability at least $1-\delta$.*

Proof. We would like to bound the difference $|J(\boldsymbol{\theta}) - \widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')|$ for a policy $\boldsymbol{\theta}' \in \Theta_\sigma$. By the definition of σ-compression, there exists at least a policy $\boldsymbol{\theta}' \in \Theta_\sigma$ such that $D_2(d^{sa}_{\boldsymbol{\theta}}||d^{sa}_{\boldsymbol{\theta}'}) \leq \sigma$. Since the IS estimator $\widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')$ is unbiased, and $\mathbb{V}\mathrm{ar}_{(s,a)\sim d^{sa}_{\boldsymbol{\theta}'}}\left[\widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')\right] < \infty$ through Lemma A.1, we can use the Chebichev's inequality to write, $\forall \epsilon > 0$,

$$Pr(|J(\boldsymbol{\theta}) - \widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')| \geq \epsilon) \leq \frac{\mathbb{V}\mathrm{ar}_{(s,a)\sim d^{sa}_{\boldsymbol{\theta}'}}\left[\widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')\right]}{\epsilon^2}.$$

Then, by calling $\delta = \frac{\mathbb{V}\mathrm{ar}_{(s,a)\sim d^{sa}_{\boldsymbol{\theta}'}}\left[\widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')\right]}{\epsilon^2}$ and considering the complimentary event, we get

$$Pr\Big(|J(\boldsymbol{\theta}) - \widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')| \leq \frac{\mathrm{R}_{\max}}{1-\gamma}\sqrt{\sigma/\delta N}\Big) \geq 1-\delta$$

where we upper bounded the variance of $\widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}')$ as in Lemma A.1 and the Rényi $D_2(d^{sa}_{\boldsymbol{\theta}}||d^{sa}_{\boldsymbol{\theta}'})$ with σ. □

Corollary 6.2. *Let Θ_σ be a σ-compression of Θ in $\mathcal{M}$ such that $|\Theta_\sigma| = K$, let $\mathcal{R}$ be a reward function for $\mathcal{M}$ uniformly bounded by $\mathrm{R}_{\max}$, let $\boldsymbol{\theta} \in \Theta$ be a target policy, and let $\delta \in (0, 1)$ be a confidence. Given N_k samples from each $\boldsymbol{\theta}_k \in \Theta_\sigma$, the error of the multiple importance sampling evaluation of $J(\boldsymbol{\theta})$ in $\mathcal{M}^{\mathcal{R}}$, i.e.,*

$$\begin{aligned}&\widehat{J}_{MIS}(\boldsymbol{\theta}/\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K) = \\ &\frac{1}{(1-\gamma)}\sum_{k=1}^K\sum_{n=1}^{N_k}\frac{d^{sa}_{\boldsymbol{\theta}}(s_{n,k}, a_{n,k})}{\sum_{j=1}^K N_j d^{sa}_{\boldsymbol{\theta}_j}(s_{n,k}, a_{n,k})}\mathcal{R}(s_{n,k}, a_{n,k}),\end{aligned}$$

is upper bounded as $|J(\boldsymbol{\theta}) - \widehat{J}_{MIS}(\boldsymbol{\theta}/\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)| \leq \frac{\mathrm{R}_{\max}}{1-\gamma}\sqrt{D_2(d^{sa}_{\boldsymbol{\theta}}||\Phi)/\delta N}$ with probability at least $1-\delta$, where $N = \sum_{k=1}^K N_k$ is a number of samples and $\Phi = \sum_{k=1}^K \frac{N_k}{N} d^{sa}_{\boldsymbol{\theta}_k}$ is a finite mixture.

Proof. Through the combination of (Papini et al., 2019, Lemma 1) and Lemma A.1, it is straightforward to derive

$$\underset{(s,a)\sim d^{sa}_{\boldsymbol{\theta}_k}}{\mathbb{V}\mathrm{ar}} \left[\widehat{J}_{MIS}(\boldsymbol{\theta}/\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K) \right] \leq \frac{(\mathrm{R}_{\max})^2 D_2(d^{sa}_{\boldsymbol{\theta}} || \Phi)}{(1-\gamma)^2 \, N}. \tag{8}$$

Then, similarly as in Theorem 6.1, we can use the Chebichev's inequality to write, $\forall \epsilon > 0$,

$$Pr(|J(\boldsymbol{\theta}) - \widehat{J}_{MIS}(\boldsymbol{\theta}/\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)| \geq \epsilon) \leq \frac{\mathbb{V}\mathrm{ar}_{(s,a)\sim d^{sa}_{\boldsymbol{\theta}_k}} \left[\widehat{J}_{MIS}(\boldsymbol{\theta}/\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K) \right]}{\epsilon^2}.$$

By calling $\delta = \frac{\mathbb{V}\mathrm{ar}_{(s,a)\sim d^{sa}_{\boldsymbol{\theta}_k}} \left[\widehat{J}_{MIS}(\boldsymbol{\theta}/\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K) \right]}{\epsilon^2}$ and considering the complimentary event, we get

$$Pr\Big(|J(\boldsymbol{\theta}) - \widehat{J}_{MIS}(\boldsymbol{\theta}/\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)| \leq \frac{\mathrm{R}_{\max}}{1-\gamma} \sqrt{\frac{D_2(d^{sa}_{\boldsymbol{\theta}} || \Phi)}{\delta N}} \Big) \geq 1 - \delta$$

where we upper bounded the variance of $\widehat{J}_{MIS}(\boldsymbol{\theta}/\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$ as in (8). □

Theorem 6.3 (Policy Optimization in Θ_σ). *Let Θ_σ be a σ-compression of Θ in $\mathcal{M}$, and let $\mathcal{R}$ be a reward function for $\mathcal{M}$ uniformly bounded by $\mathrm{R}_{\max}$. The policy $\boldsymbol{\theta}^*_\sigma \in \arg\max_{\boldsymbol{\theta} \in \Theta_\sigma} J(\boldsymbol{\theta})$ is ϵ-optimal for the MDP $\mathcal{M}^{\mathcal{R}}$, where $\epsilon := |\max_{\boldsymbol{\theta}\in\Theta} J(\boldsymbol{\theta}) - J(\boldsymbol{\theta}^*_\sigma)| \leq \frac{\mathrm{R}_{\max}}{1-\gamma} \sqrt{\log \sigma}$.*

Proof. Let be $\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta}\in\Theta} J(\boldsymbol{\theta})$. From the definition of σ-compression we have that there exists at least a policy $\boldsymbol{\theta}' \in \Theta_\sigma$ such that $D_2(d^{sa}_{\boldsymbol{\theta}^*} || d^{sa}_{\boldsymbol{\theta}'}) \leq \sigma$. Then, we can write

$$\begin{aligned}
(1-\gamma)|J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}')| &= \left| \int_{\mathcal{SA}} \mathcal{R}(s,a) \big(d^{sa}_{\boldsymbol{\theta}^*} - d^{sa}_{\boldsymbol{\theta}'} \big) \, \mathrm{d}s \, \mathrm{d}a \right| && (9) \\
&\leq \mathrm{R}_{\max} \int_{\mathcal{SA}} \left| d^{sa}_{\boldsymbol{\theta}^*} - d^{sa}_{\boldsymbol{\theta}'} \right| \mathrm{d}s \, \mathrm{d}a && (10) \\
&\leq \mathrm{R}_{\max} \sqrt{d_{KL}(d^{sa}_{\boldsymbol{\theta}^*} || d^{sa}_{\boldsymbol{\theta}'})} && (11) \\
&\leq \mathrm{R}_{\max} \sqrt{\log \big(D_2(d^{sa}_{\boldsymbol{\theta}^*} || d^{sa}_{\boldsymbol{\theta}'}) \big)} = \mathrm{R}_{\max} \sqrt{\log \sigma} && (12)
\end{aligned}$$

where (9) is from the definition of J given in Section 2.2, (11) is obtained from (10) through the Pinsker's inequality, and (12) derives from $d_{KL}(p||q) = d_1(p||q) \leq d_2(p||q) = D_2(p||q)$, which is straightforward from the definition of Rényi divergence. Finally, it is trivial to see that $J(\boldsymbol{\theta}^*_\sigma) \geq J(\boldsymbol{\theta}')$ for $\boldsymbol{\theta}^*_\sigma \in \arg\max_{\boldsymbol{\theta}\in\Theta_\sigma} J(\boldsymbol{\theta})$. □

Theorem 6.4 (Off-Policy Optimization in Θ). *Let Θ_σ be a σ-compression of Θ in $\mathcal{M}$ such that $|\Theta_\sigma| = K$, let $\mathcal{R}$ be a reward function for $\mathcal{M}$ uniformly bounded by $\mathrm{R}_{\max}$, and let $\delta \in (0,1)$ be a confidence. Given N_k samples from each $\boldsymbol{\theta}_k \in \Theta_\sigma$, we can recover an ϵ-optimal policy for $\mathcal{M}^{\mathcal{R}}$ as*

$$\begin{aligned}
\big(\_ \, , \, \boldsymbol{\theta}^*_{IS} \big) \in \underset{\boldsymbol{\theta}_k \in \Theta_\sigma, \boldsymbol{\theta}\in\Theta : D_2(d^{sa}_{\boldsymbol{\theta}} || d^{sa}_{\boldsymbol{\theta}_k})}{\arg\max} \\
\frac{1}{(1-\gamma)N_k} \sum_{n=1}^{N_k} w_{\boldsymbol{\theta}/\boldsymbol{\theta}_k}(s_n, a_n) \mathcal{R}(s_n, a_n),
\end{aligned} \tag{7}$$

such that $\epsilon := \big| \max_{\boldsymbol{\theta}\in\Theta} J(\boldsymbol{\theta}) - J(\boldsymbol{\theta}^*_{IS}) \big| \leq \frac{\mathrm{R}_{\max}}{1-\gamma} \sqrt{2\sigma / N_k \delta}$ with probability at least $1 - \delta$.

Proof. Thanks to the definition of σ-compression and the guarantee provided by Theorem 6.1, from the collected samples we have that there exists $\boldsymbol{\theta}_k \in \Theta_\sigma$ such that

$$\widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}_k) - \frac{\mathrm{R}_{\max}}{1-\gamma} \sqrt{\frac{2\sigma}{N_k \delta}} \leq J(\boldsymbol{\theta}) \leq \widehat{J}_{IS}(\boldsymbol{\theta}/\boldsymbol{\theta}_k) + \frac{\mathrm{R}_{\max}}{1-\gamma} \sqrt{\frac{2\sigma}{N_k \delta}}$$

holds $\forall \boldsymbol{\theta} \in \Theta$ with probability at least $1 - \delta/2$. Then, let $\boldsymbol{\theta}^*_{IS}$ be a policy obtained as in (7), and let $\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta})$. We consider the event in which $J(\boldsymbol{\theta}^*_{IS})$ falls below its lower confidence bound and $J(\boldsymbol{\theta}^*)$ exceeds its upper confidence bound. It is easy to see that this event happens with probability at most δ, whereas the complimentary event guarantees that

$$|J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}^*_{IS})| \leq \frac{\mathrm{R}_{\max}}{1-\gamma} \sqrt{\frac{2\sigma}{N_k \delta}}.$$

□

B. Optimization Problems

B.1. Quadratic Program Formulation of (5)

The optimization problem in (5) can be formulated into a quadratically constrained quadratic program as

$$\begin{aligned}
\underset{z \in \mathbb{R}, \boldsymbol{\omega} \in \mathbb{R}^{\mathcal{SA}}}{\text{maximize}} \quad & z \\
\text{subject to} \quad & z - \int_{\mathcal{SA}} \frac{\big(\omega(s,a)\big)^2}{d^{sa}_{\boldsymbol{\theta}^*_k}(s,a)} \,\mathrm{d}s\,\mathrm{d}a \leq 0, \qquad \forall k \in [K] \\
& \int_{\mathcal{A}} \omega(s,a)\,\mathrm{d}a = (1-\gamma)\mu(s) + \gamma \int_{\mathcal{SA}} \omega(s',a') P(s|s',a')\,\mathrm{d}s'\,\mathrm{d}a', \qquad \forall s \in \mathcal{S} \\
& \omega(s,a) \geq 0, \qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}.
\end{aligned}$$

B.2. Linear Program Formulation of (6)

The optimization problem in (6) can be formulated into a linear program as

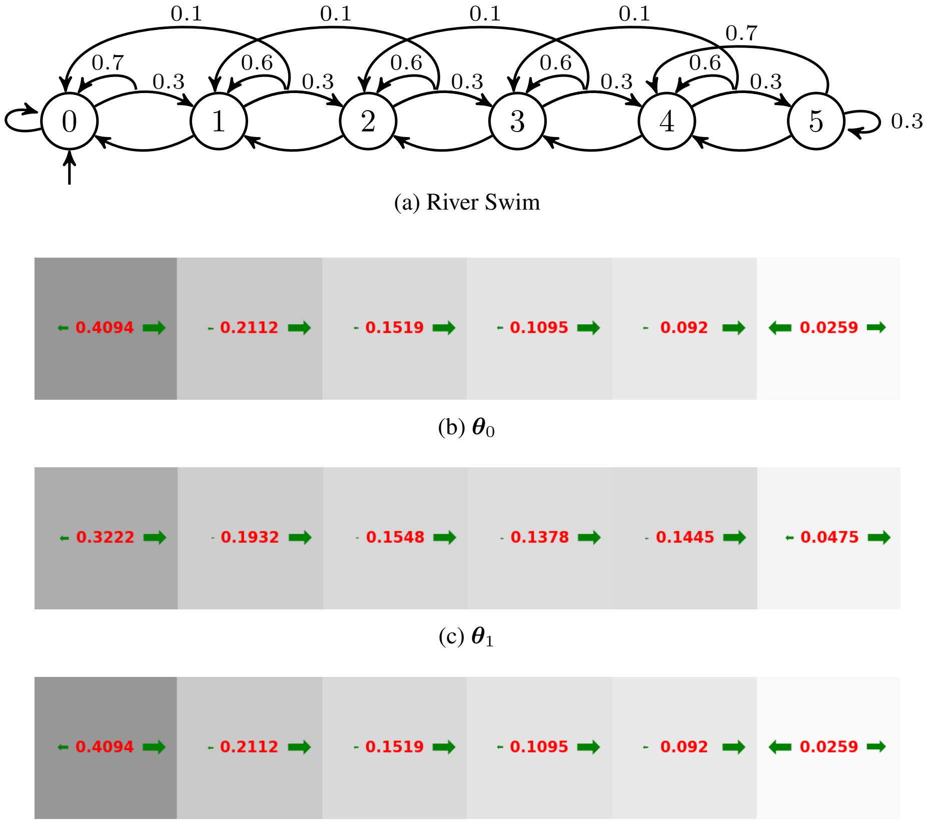
$$\begin{aligned}
\underset{z \in \mathbb{R}, \boldsymbol{\omega} \in \mathbb{R}^{\mathcal{SA}}}{\text{maximize}} \quad & z \\
\text{subject to} \quad & z - \int_{\mathcal{SA}} \frac{\omega(s,a)}{d^{sa}_{\boldsymbol{\theta}^*_k}(s,a)} \,\mathrm{d}s\,\mathrm{d}a \leq 0, \qquad \forall k \in [K] \\
& \int_{\mathcal{A}} \omega(s,a)\,\mathrm{d}a = (1-\gamma)\mu(s) + \gamma \int_{\mathcal{SA}} \omega(s',a') P(s|s',a')\,\mathrm{d}s'\,\mathrm{d}a', \qquad \forall s \in \mathcal{S} \\
& \omega(s,a) \geq 0, \qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}.
\end{aligned}$$

C. Further Details on the Numerical Validation

In Section 7.1, we commented the results of PSCA in the River Swim domain. For the sake of clarity, here we report an illustration of the River Swim CMP (Figure 2a), heatmap visualizations of the policies in the σ-compression obtained by PSCA (Figure 2b-2d), and the set of parameters we employed ($\sigma = 10, \alpha = 0.005, \beta = 0.1$). We further report the results of an additional policy space compression experiment in a Gridworld domain ($|\mathcal{S}| = 9, |\mathcal{A}| = 4$). In this setting, we considered $\sigma = 40, \alpha = 0.005, \beta = 0.1$, and the resulting σ-compression is composed of $K = 4$ policies (a visualization is provided in Figure 3a-3d).

In Section 7.2, we reported a set of policy evaluation experiments in the River Swim domain. Especially, we considered an IS off-policy evaluation setting, in which we take a batch of samples with each policy $\boldsymbol{\theta}_k \in \Theta_\sigma$, or with a uniform policy $\boldsymbol{\theta}_{\mathcal{U}}$, or with the target policy itself $\boldsymbol{\theta}$. For every policy, the batch is composed of $N = 100000$ samples, and it is obtained by drawing 5000 trajectories of 20 steps. Similarly, we considered a MIS off-policy evaluation setting, in which we take a batch of samples with the σ-compression Θ_σ, or a set of three random policies Θ_3. In both the cases, the batch is composed of $N = 300000$ samples ($N_k = 100000$ for each policy in the space), obtained by drawing 15000 trajectories of 20 steps.

In Section 7.3, we reported a policy optimization experiment in the River Swim domain. To run this experiment, we implemented the action-based formulation of the OPTIMIST algorithm (Papini et al., 2019, Algorithm 1). For each seed, we run the algorithm for 100 iterations, in each iteration we collect $N = 1000$ samples, which are obtained from 50 trajectories of 20 steps. The value of the importance weights truncation M and the confidence schedule δ_t are taken from the theoretical analysis in (Papini et al., 2019).



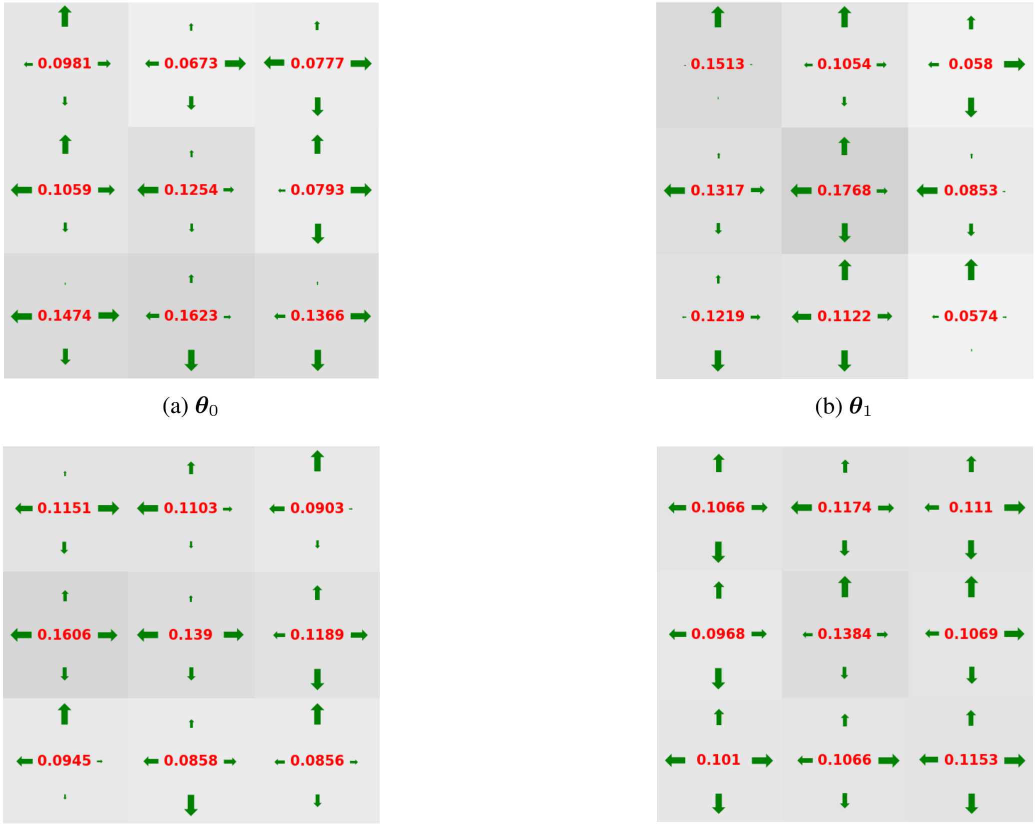
(a) River Swim

(b) $\boldsymbol{\theta}_0$

(c) $\boldsymbol{\theta}_1$

(d) $\boldsymbol{\theta}_2$

Figure 2. **(a)** Illustration of the River Swim CMP. **(b, c, d)** Heatmap visualization of the policies in the σ-compression $\boldsymbol{\theta}_k \in \Theta_\sigma$ for the River Swim domain. The background color and the label denote the state probability, the green arrows represent the policy in the state.



(a) $\boldsymbol{\theta}_0$

(b) $\boldsymbol{\theta}_1$

(c) $\boldsymbol{\theta}_2$

(d) $\boldsymbol{\theta}_3$

Figure 3. **(a, b, c, d)** Heatmap visualization of the policies in the σ-compression $\boldsymbol{\theta}_k \in \Theta_\sigma$ for the Gridworld domain. The background color and the label denote the state probability, the green arrows represent the policy in the state.